

**DUE DATE SLIP****GOVT. COLLEGE, LIBRARY****KOTA (Raj.)**

Students can retain library books only for two weeks at the most.

BORROWER'S No.	DUE DTATE	SIGNATURE

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE  
STUDIES IN STATISTICS  
AND SCIENTIFIC METHOD

*Edited by*

PROFESSOR A. L. BOWLEY AND PROFESSOR A. WOLF

No. I

ELEMENTARY  
STATISTICAL METHODS

# ELEMENTARY STATISTICAL METHODS

By

E. C. RHODES, B.A., D.Sc.

*Reader in Statistics in the University of London*

With 38 diagrams

LONDON

GEORGE ROUTLEDGE & SONS, LIMITED

BROADWAY HOUSE, 68-74 CARTER LANE, E.C. 4

*First published*      November 1933  
*Second impression*   September 1935  
*Third impression*      August 1937  
*Fourth impression*   January 1941  
*Fifth impression*      January 1944  
*Sixth impression*      May 1945  
*Seventh impression*   July 1946  
*Eighth impression*   January 1948

PRINTED IN GREAT BRITAIN BY  
 LIND HUMPHRIS  
 LONDON      BRADFORD

# CONTENTS

CHAP.	PAGE
1. STATISTICS . . . . .	I
2. STATISTICAL INQUIRIES . . . . .	7
3. ASSEMBLING STATISTICAL DATA . . . . .	22
4. SECONDARY OR DERIVED STATISTICS . . . . .	42
5. COMPARISON OF AVERAGES . . . . .	60
6. THE CALCULATION OF AVERAGES . . . . .	77
7. GRAPHICAL METHODS . . . . .	97
8. THE MEDIAN AND MEASURES OF DISPERSION	117
9. WEIGHTED SUMS AND WEIGHTED AVERAGES .	136
10. INDEX NUMBERS . . . . .	151
11. GRAPHS OF TIME SERIES . . . . .	173
12. ANALYSIS OF TIME SERIES . . . . .	208
INDEX . . . . .	240

# ELEMENTARY STATISTICAL METHODS

## CHAPTER I STATISTICS

The functions of a Statistician may properly be considered as divisible into three parts. In the first place he is concerned with the assembling of statistical data, in the second place with their analysis, in the third place with the interpretation of the results of such an analysis. Sometimes, owing to the division of labour in a highly organized society, a particular statistician may only be concerned with one or two of these functions. Thus, he may be solely engaged in the analysis of statistics, without bothering himself about the methods of collection of the data or with the possible interpretations which may be put to his results. This kind of subdivision of the statistician's duties may not always be favourable to the best elucidation of a particular problem, for the persons assembling the data may know little or nothing of the superstructure of deduction which is subsequently to be built on the foundation of facts—if they knew more they might modify somewhat their methods, in order that this superstructure would have less likelihood of being blown down by a storm of criticism. Similarly, those

who attempt only the interpretation of the results of a statistical analysis, without knowing much of the sources of the data and the methods of analysis, would perhaps modify their conclusions if they had followed through the whole process from the raw material stage to that of finished product.

At whatever stage a statistician is concerned with the data, he ought to know something definite about the processes to which these data are submitted in their earlier and later phases of treatment. Otherwise we get the phenomenon of persons earnestly engaged in analysing statistical data without knowing exactly why they are doing so, or persons collecting data which are valueless in their present form, but which might have been of great use if the method of obtaining them had been slightly modified, or others putting an interpretation on the results of an analysis which the original data do not justify.

We must therefore concern ourselves with these various aspects of the work of a statistician, and in the first place we must ask ourselves "What are statistical data?" These in their original form are facts relating to a group of units which are susceptible of counting or numerical appreciation in some form or other. Thus, we may be concerned with a number of factories manufacturing a given product. The numerical facts relating to machines in operation, persons employed, hours worked, wages paid, and so on are statistical data. We may be dealing with farms, their acreage, the number of livestock, and so on. We may be dealing with heights of a group of persons or their incomes or their state of health. Statistical data are available in large quantity in records

of various kinds, books kept by firms, publications relating to trade, population, and so on.

Such pieces of information, in order to be appreciated at their true worth, should be precise. Considerations of space and time are elementary as the foundations of such precision. We must know definitely to what particular region or place the statistics refer, and to what instant or period of time they have relation. This regard to space and time is but one aspect of a general consideration of exactness of definition of particular terms which are used in the description of the statistics. For a proper appreciation of the exact meaning of statistical data, we must be acquainted with the precise significance of such words as "population", "manufacturing concern", "trade", "port", etc., in the context. All such definitions necessarily involve reference to an area and to time.

We realize, immediately we think about the matter, that statistical data are only of value for comparative purposes, and, realizing this, we see the necessity for precision, because if we wish to compare two things which are described in the same way on different occasions, we must be sure that the same words used on these occasions have really the same meaning.

We may, on occasion, make a vague statement such as, "The population of England and Wales nowadays is very large", and such a statement in its right context probably conveys quite adequately our intention. Or we may say instead "The population of England and Wales nowadays is about 40 millions"; here we replace the adjectival description "very large" by "about 40 millions" and probably convey the same kind of

impression to others; we might equally well have used "45 millions" or "50 millions" to convey the same impression of the large size of the population, and no one would worry to examine the exactitude of the figure. No precise significance attaches to such statements as "There are 3,000,000 people in Great Britain suffering from nervous disorders"—Dr. Elizabeth Sloan Chesser, *Evening Standard*, 4th April, 1933. The point is that in such bare statements of fact the figures are really being used in an adjectival sense. But when we come to other statements involving comparisons, we have to make certain that our figures are correct. We may be comparing the value of Imports into the United Kingdom in 1913 with those in 1931. Now we have to be sure, (1) that we are dealing with the same kind of goods, i.e. that the definition of "Imports" has not changed between the two dates, (2) that we are dealing with the same geographical area, i.e. that the term "United Kingdom" means the same in 1931 as it did in 1913 (as a matter of fact this is not the case owing to the separation of the South Ireland trade figures from the United Kingdom figures on the formation of the Irish Free State in 1923), (3) that the figures themselves are accurate. Similarly we may desire to review the position as regards Unemployment in 1931 in Great Britain, the United States, and Germany by comparing the Unemployment Statistics of these three countries. Before we can come to correct conclusions from such an examination of these figures, we must consider whether the same kinds of definitions are used in these countries of the term "Unemployment Statistics"; (as a matter of fact grave difficulties are encountered in attempts at international statistical

comparisons owing to different interpretations being given, for administrative or statistical purposes, to simple terms which, in the opinion of the man in the street, may only have one meaning).

It is essential to realize, from the start, that in certain cases the statistical aspect may be subsidiary to other considerations. For instance, it may be desirable to compare the state of civilization of a country with that of another, or of one country at different epochs. Such a comparison may be attempted by a description of summaries of observers' experiences, or recourse may be had to such figures as are available, which appear to have relation to the particular problem, such figures as the numbers of persons obtaining School Certificates each year, or figures giving the numbers of places of worship or places of entertainment, or figures giving numbers of persons convicted of crime, and so on. The "state of civilization" is not susceptible of statistical treatment. We may be pleased to utilize these sort of figures as of secondary importance, but major consideration would be given to general descriptions of the way people live, the kind of work they do, how they spend their leisure time, what they eat, and so on.

Similarly in inquiries relating to the social conditions of the population, housing statistics, giving, for instance, the proportions of houses where overcrowding is deemed to exist (a statistical measure of persons per room having been achieved and passed), poverty statistics, giving proportions of population relieved by the Public Authority, and such like, have to be related to other facts, such as the type of houses, type of room in the house, location of the home in relation to the workplace, location of the

family near relatives, and *so on*, which may not easily be susceptible of statistical appreciation. Therefore, when we are considering a particular problem, the statistics must be given their proper weight, and no more than this, in relation to the other factors which are involved and which cannot be dealt with statistically.

## CHAPTER 2

### STATISTICAL INQUIRIES

The phrase "Statistical Inquiries" is used in the present instance to cover a large field of action. Statistical data emerge in a variety of ways. They may be produced as a by-product of certain administrative operations. For instance, Imports into the United Kingdom, which are subject to duty, are carefully checked on entering by the Customs Officials, because the law requires that levies should be made on these classes of goods entering into consumption in the country. A necessary feature of this kind of action is the keeping of records of such entries. These records serve as the raw material of Statistical Tables, which are later available to the public, giving information on the subject of this class of Imports. Similarly, there are laws relating to 'Unemployment Insurance which have to be administered. In the process of administration records are kept, and these records again represent raw material of statistics. In such cases as these the collection of statistical data is not the primary purpose; it is doubtful if the necessary action would be taken if the collection of statistics were all that was intended. On the other hand such statistical data are immensely valuable, because they give us definite information on certain matters of importance in the National Economy.

But statistical data are obtained directly as the result of an effort made to get information about certain affairs

of moment. For instance the Census of Population results in statistics which it is the purpose of the Census to obtain. In like manner the Census of Production gives statistical data relating to the Output of Industry, in a particular period of time, and to other cognate matters, and this information is obtained during the course of an inquiry, the sole purpose of which is the collection of these data.

For present purposes it is useful to consider that the words "Statistical Inquiries" refer to both these types of inquiry, the first where the collection of statistics is subsidiary to or a part only of the main action, the second where the collection of statistics is the only end in view. Appreciation of these two kinds of Statistical Inquiry is necessary because one of the fundamental factors to be considered in handling statistics is their accuracy and their meaning. Now, the meaning of statistical data depends upon the precise definitions given to certain terms to which the figures relate, which may often be technical in character, and these definitions may be primarily made to assist certain administrative action, or they may be made to suit a particular statistical inquiry which is undertaken with the idea of obtaining information relating to certain matters of interest. For instance, the word "Imports" connotes a certain meaning in the mind of the ordinary person, he thinks vaguely of all goods coming into the country from abroad, but "Imports" for purposes of understanding the Statistical Tables issued by the Board of Trade means something quite precise and unambiguous.<sup>1</sup> Similarly the words "Unemployed Person" convey a vague impression to the man in the street, but for the purpose of administering the Insurance

<sup>1</sup> See Appendix to Chapter 2.

Acts the Ministry of Labour have devised a workable definition which assists them in their operations, and which resolves the vagueness of the impression which the ordinary person has when he thinks of "Unemployed Persons".<sup>1</sup> Similarly in those investigations conducted by Professor Bowley and others into Working-Class Conditions of living in certain towns in 1913-14, 1924-5, the results of which were published in "Livelihood and Poverty" and "Has Poverty Diminished?", and in the New Survey of London Life and Labour, 1928, an appropriate definition was invented for the term "Working-Class".<sup>1</sup> The difference between the two types of inquiry may be considered as one which influences the definitions given to terms used; where the statistical data emerge as subsidiary to the main action the definitions may be made primarily to suit administration, and where the statistics emerge as the sole function of the inquiry the definitions will be made primarily to suit the conditions of the problem on which the statistical data will throw light. This emphasis on the meaning of statistical data, which involves knowledge of definitions used, is necessary on account of the fact that statistics only serve a useful purpose when placed in relation to one another for comparisons to be made. So we need to know, if we are dealing with spatial comparisons, that the definitions used are the same in two countries, for instance; or, if dealing with comparisons involving considerations of time, that the definitions used are the same at one time as at another. Now, even if two countries have the same kind of laws, their administration of them may be different, consequently the meaning of statistical data which emerge as a result may be different. Again, a country may modify

<sup>1</sup> See Appendix to Chapter 2.

a particular series of laws from time to time, thus the administration may alter and the meaning of statistical data may be changed, although the same terms be used to describe these data. On the other hand, if an inquiry has been instituted with the object of ascertaining certain facts, and as a consequence particular definitions are invented, similar inquiries in other countries or at other times may be instituted with the first as a model and the same definitions used again.

With regard to the subject of the accuracy of the statistical data, it is obvious that sufficient precision is required in the results of an investigation, if the information obtained is to have any weight in subsequent argument or discussion. The accuracy of a final result depends on the accuracy of each part which contributes to a total. It is reasonable to suppose that less care will be taken in ensuring accuracy in statements and figures, if the primary use of such is not statistical but (say) administrative, where perhaps rough approximations to the truth are as useful as closer approximations. It is not meant that wilful misrepresentations are made deliberately, but merely that, in the stress and turmoil of doing a job, a person may not have the time or inclination to check every statement and figure, which are accepted at their face value, so long as they do not indicate any great apparent divergence from what is likely to be correct. For instance, when goods are entered into the United Kingdom as Free Imports, the Importers render statements to the Customs House which are checked, but for the most part they are assumed correct without further inquiry. In the case of Dutiable Goods Imported, more scrutiny is of course required. If the sole purpose

of the Custom House Officials were the collection of statistics a much larger staff would be required in order that all statements rendered could be properly checked and absolute accuracy insisted on. Naturally, if the purpose of an inquiry is merely the search for information it is more likely that the results will be reasonably accurate. In general, therefore, we may suppose that the information obtained as a result of the first kind of statistical inquiry is not likely to be as accurate as that obtained as a result of an inquiry of the second kind, that is, an inquiry *ad hoc*.

The subject of accuracy brings to mind another distinction between statistical inquiries. This distinction depends on the source from which the information is obtained. In some inquiries a large number of persons play an influential part, in others a comparatively small number are actively involved, and those are selected in some way. This distinction is also related to the scope of the inquiry. In the Census of Population, for instance, each householder is responsible for the information relating to the persons in his household and for details as to the domicile. The number of householders is very large indeed, and therefore the sources from which the information is obtained in the Census of Population are many and varied. There are such wide variations in the educational standards and intelligence of the householders in this country, that great care has to be taken in the wording of the questions put on the official form in order to avoid the possibility of ambiguity, and even then, if the questions are understood, some people may find themselves in difficulties over the answering of them. Moreover, this wide range of variation in the sources from which the

information is obtained is one reason why the Census Office only attempts to get knowledge of a comparatively small number of matters of interest. The scope of the inquiry is necessarily limited on this account. Even with the small number of questions asked in the Census, doubt must arise as to whether the questions have been answered truthfully or not, since no check is possible, except in the most obvious cases. There may be no particular incentive to supply wrong information, it may be the fact that sheer lack of accurate knowledge forces a person unwittingly to answer a question wrongly. We conclude that the accuracy of the final figures, depending on the accuracy of the constituent parts, is dependent on the number of sources from which the information is obtained.

On the other hand these sources may be comparatively few in relation to the size of the inquiry. These few may be skilled investigators whose duty it is to cover the ground collecting the information. These persons may themselves examine others from whom the information is received, but by judicious questioning and cross questioning, or by reference to other sources of information, they may easily assure themselves of the truth of the information collected. Moreover, if adjectival descriptions, rather than accurate measurements, are given in answer to certain questions, it is more likely that these investigators would maintain a reasonable standard of value attaching to certain descriptive terms, than that the same meaning would be attached to these terms by a host of individuals. For instance, School Medical Inspections are made regularly, where school children are examined by Medical Officers in the service of the Educational Authority, and at these examinations

reports are made on the children's height, weight, state of nurture, condition of teeth, throat, hair, and so on. To a certain extent such words as Normal, Good, Poor are used in this connection, and it is obviously better that such descriptions should be given by the few doctors rather than by the many children or parents. Otherwise no one would be sure that the information obtained was of any use. Further, it is likely that the scope of an inquiry can be extended by using investigators who survey the field of inquiry, because they are often able to elicit information on particular topics, by subtly varying the kind of question put, in cases where the usual form of question is not properly understood, and by impressing the need for such an inquiry on the mind of the person from whom the information is desired. Such information would perhaps not be obtained from many persons if they were solicited by means of a questionnaire, schedule, or official form, either because the persons do not understand what is being asked, or how to answer, or because they do not see any real reason why they should answer. This method has been used with success by Professor Bowley and his collaborators in the investigations referred to previously and has elicited much valuable information about details of family life in working-class households.

A further distinction must be made between different kinds of statistical inquiries. The information obtained during the course of an investigation concerns a number of things, animate or inanimate, and the investigation is circumscribed by a set of limiting conditions which determine its scope, or the extent of the survey. We may say, in brief, that the information sought is that relating to a group of units, whether these units are human beings,

or cattle, or farms, or crops, or manufacturing concerns, or machines, or ships or consignments of goods does not matter. The problem is to extract useful information about the group. Here comes the distinction between different kinds of inquiry. The investigation may be concerned with the whole of the group or with only a part of it. These two kinds of inquiry may be differentiated by using the words *Census* and *Sample*. In the *Census*, the whole field is surveyed, in the *Sample* only a part of the field is surveyed. From a census we consider that we have obtained facts relating to the whole number of units coming within the scope of the inquiry, and are under no apprehensions when we seek to make deductions in general terms on the problem before us. On the other hand, when we have only a sample inquiry, our deductions from the facts assembled only relate to this sample and, if we wish to generalize in terms of the whole group sampled, we must be sure that our sample is a representative one. Illustrations of the census type of inquiry are the *Census of Population*, the *Census of Production*, the *Accounts of Import and Export Trade*. In the *Census of Population* information is desired of all the persons who form part of the population, and every endeavour is made to ensure that all these units are brought within the scope of the inquiry. In the *Census of Production 1930*, information was required from all manufacturing concerns employing more than 10 workpeople, but the *Census of 1924* was extended to all firms, however small. In the *Accounts of Trade* information is given concerning all consignments of goods entering or leaving the country, coming within the definitions of *Imports and Exports*. Naturally with such large-scale investigations as these, much expenditure of time and labour is necessary before the end is reached.

Both these considerations imply expense. There are three factors to be taken into account, (1) the desirability of obtaining information, (2) the necessity for this information to be available to those who would use it within a reasonable time after the date when the information is collected, (3) the expense involved in obtaining, collating, and presenting this information. In the case of the Census of Population the information has only been obtained every ten years, and the data and conclusions are only available some considerable time after the period to which the information relates. (The final reports relating to the 1921 Census in Great Britain and Northern Ireland were published in 1927.) A Census of Production was made in respect of 1907, 1912, 1924, 1930. The final tables relating to 1924 are now being issued by the Board of Trade (1932).<sup>1</sup> On the other hand, information relating to the Foreign Trade is available month by month in considerable detail within a fortnight after the end of the month to which the data relate, and full details of the trade of a particular year are published within two years. This is only possible by using a large staff of Custom Officials.

At the same time as the Census of Production Inquiry 1924 the concerns which were required by law to render information respecting their products, numbers employed, power used, etc., were circularized also by the Board of Trade, to the effect that they should voluntarily submit statements of wages paid, and hours worked by work-people, etc., during the same period. Not all, but considerable proportions, responded to this invitation and

<sup>1</sup> The preliminary reports on the 1930 Census of Production have been available as supplements to the Board of Trade Journal in 1932 and in the first part of 1933

there resulted a large amount of useful information on the subject of earnings and hours worked, relating to this sample of Industry. It was felt that this sample was sufficiently representative of the whole so that the facts which emerged might be considered as applicable to Industry as a whole. This is an illustration of what is meant by a sample inquiry.

There are two kinds of sample inquiry. The one where the investigators have no control over the formation of the sample, where the net is cast and satisfaction is felt with the result of the fishing, the other where the investigators choose deliberately the particular units which are allowed to form part of the sample, and every endeavour is made to obtain information about those, and only those. This second method is, of course, only possible when the limitations of the whole group, of which a sample is taken, are clearly defined and where such picking and choosing can be done. In this case an effort is made to choose a random sample, which from the start will, it is hoped, be representative. The method is to pick the individual units out of the whole group by some mechanical process, which allows every unit the same probability of entering the sample and where blind chance alone determines whether one or another is chosen. This method was used by Professor Bowley and associates in the investigations referred to previously. The problem was the finding of information relating to working-class conditions of living in a number of towns. The local directory gave for each town the limits of the whole group to be sampled. The purely mechanical process of turning over the pages of this directory, marking there each twentieth address as it occurred gave a twentieth

sample of the whole town. Information was sought about the families at the marked addresses. This same method has been used in recent years by the Ministry of Labour in order to get more detailed information than had been available previously, respecting that part of the population coming within the scope of the Unemployment Insurance Acts, to which the unemployed belong. Registers of names are kept by the Central Authority, these were consulted, names picked out at definite intervals, and those persons assumed to form the sample. Details were required of these persons only. This method of choosing the constituent parts of the sample gives what is hoped to be a random sample of the total group.

In the first method the investigators have no control over the sample and do not know, when the sample is obtained, whether it is representative or not. For instance, an Insurance Company doing life business attracts a certain number of members into insurance; from each, certain particulars as to age, family history, occupation, and so on are obtained. At any given time those entering into contracts with the Company within the past year, say, may be considered as a sample of the Insuring Class of the whole population, and, if the Company knows of no reason why there should be constraining forces at work which would tend to make theirs a biased sample, this sample is a random one of this class of the population. But the onus of proof is on the Company; it cannot be assumed off-hand that the sample is random, and if any conclusions were to be drawn from this sample which should be assumed equally applicable to the whole of this class of the population, then somehow it must be demonstrated that the sample does appear to be

representative. This question of the random nature of the sample must be dealt with before generalizations are permitted

There are certain statistical tests which may be used to determine whether the sample is representative or not, and these tests should be applied in all cases of sampling, even where the sample has been obtained by mechanical choosing of those who form part of it, that is, in the case of those samples obtained by the second method described above. This is necessary because we are not certain—

- (a) that the method of choosing has been operated correctly,
- (b) that there was not some biased order or arrangement of the units in the whole group from which the sample is chosen, and that this bias may not have prejudiced the random nature of the sample.

With regard to the samples obtained by the first method, that is, when the investigators have no control, these tests are certainly necessary; they supply the only evidence of the random nature of the sample, if it is of this kind. These tests will be described later. (See Appendix to Chapter 4)

The *Sample* method of Inquiry has many advantages over the *Census* method. The expense of time and labour involved in the *Census* method are no longer in evidence. With a comparatively small number of units in the *Sample*, the assembly and analysis of the statistics is reduced considerably, much time is saved, and much labour; the results of the inquiry are available reasonably soon after the inquiry is instituted. Moreover, in many cases,

with a sample inquiry, if the sample is not too large, skilled investigators can be used for the collection of the data, and instead of a vast multitude of sources from which the information is obtained, with a consequent possible lack of accuracy, there is a small number of sources. Where a random sample, obtained by the second method described above, is possible, it is certainly preferable to the Census method, and its use is becoming more extensive.

## APPENDIX

### DEFINITION OF IMPORTS

Quoted from the Monthly Accounts relating to Trade and Navigation of the United Kingdom.

"The particulars in respect of imported goods from which the official trade statistics are compiled are allowed to be given by Importers or their Agents at any time within fourteen days after the arrival of the ship. Further extension of time is given within which to make any necessary amendments. . . . It follows that the statistics published for a month do not precisely represent the imports . . . which occurred in that period. . . . The following classes of goods arriving in this country are not included in the Import Statistics:—

(a) Personal luggage, including parcels brought by passengers for private use, so long as such parcels do not contain dutiable goods. Dutiable goods contained in passengers parcels are included in the statistics ;

(b) Fresh fish and shell fish of British taking, landed from British ships arriving direct from the fishing grounds ;

(c) Ships' stores, military and naval stores on board

government vessels, bunkers (coal and oil), and ballast of no commercial value ;

(d) Mats, sacks, cases, etc., used as packages of imported goods ;

(e) Goods directly imported by Ambassadors and Ministers accredited to this Kingdom ;

(f) Old vessels bought by owners from abroad.

### NUMBER OF UNEMPLOYED

Each person coming within the scope of the Unemployment Insurance Acts is supplied with a book, which is supposed to be lodged with an Employment Exchange when the person registers as unemployed. The number of Unemployed at any time is determined by the number of books so " lodged ".

### WORKING CLASS

The New Survey of London Life and Labour undertaken in 1928 took account of particulars relating to a large sample of households in London. Those details referring to Middle Class households were excluded from the subsequent analysis, so that in effect the definition of " Working Class " was a negative one. The following information is given in Volume 3, Appendix I, of the New Survey :—

" Middle Class : Middle class households are distinguished primarily by the occupation of the head. But some measure of discretion must be employed.

## Special Cases :

(1) Professional and clerical occupations to be ranked middle class. This includes commercial travellers, insurance agents, etc.

(2) All publicans to be ranked middle class.

(3) Shopkeepers to be ranked middle class unless the shop is subsidiary to the work of the principal wage-earner, or the income from the shop is definitely below £250.

(4) Self-employers, small employers, master-men, etc., not to be ranked as middle class unless their incomes are definitely over £250 a year.

(5) Hawkers, street-traders, etc., to be ranked working class.

(6) Shop assistants to be ranked working class unless their work is managerial or supervisory (e.g. departmental head or shopwalker) or unless their wage suggests middle class rank.

(7) Police sergeants to be ranked working class, inspectors middle class."

## CHAPTER 3

### ASSEMBLING STATISTICAL DATA

Having dealt with the kinds of Inquiry, we must now consider the nature of the information which has been obtained. The essential fact which must be appreciated is that whatever the inquiry is concerned with, whatever the nature of the inquiry, the information collected refers to a number of individuals or units of some kind or another, which together form the group with which the inquiry is concerned. The units may be households, persons, houses, sheep, farms, ships, consignments of goods, mines, firms, and so on. Every unit possesses a number of peculiarities, characteristics, or attributes, and the units possess these in a variety of ways or degrees, and it is the varying nature of the possession of these characteristics by different units which enables us to distinguish between them. The number of these characteristics is large, some of them are susceptible of measurement, some are merely accorded an adjectival description. For instance, a unit in a group may be a male person, aged 26 years, 5 ft 9 in. in height, 10 st. 9 lb. in weight, engaged as a clerk with a particular firm, dwelling in a certain house, having an income of £150 a year, married with one child, wearing on the 1st of July, 1931, a blue suit, green tie, black shoes, and having blue eyes, black hair, and so on. When we exhaust all this person's characteristics we isolate him from others. The sum total of his characteristics enables his friends to know him apart from others and establish his identity as a unit.

Similarly, a unit in a group may be a ship, using steam as motive power, engaged in the Foreign Trade, registered under the British flag, arriving at Hull on the 21st July, 1931, with a particular cargo, of a certain tonnage, and carrying so many in the crew, and so on. The sum total of these characteristics determine the vessel's identity. So with a manufacturing concern, it is situated in a particular place at a certain time, it is engaged in a certain trade, it employs so many hands, it uses so much horsepower, it has so many machines, it has so many storeys, so many windows, so much floor space, and so on. These illustrations indicate sufficiently what is meant by peculiarities, characteristics or attributes, and the fact that some are indicated by a number, e.g. height, and some by a description, e.g. eye-colour. Now, in a given inquiry we are generally only concerned with a number of these characteristics, we may not, for instance, be interested in persons' heights, but we may be interested in their incomes; on the other hand, in other inquiries we may be concerned with heights and not with incomes. The information collected, then, in the course of an inquiry concerns a number of units in a group, defined and limited in some way or another, and of these units we obtain details as to the possession of one or more, but probably not all, of their characteristics. The limits of the group in the inquiry are determined by the units all possessing some one or more characteristics in the same manner or degree. For instance, in the Census of Population in England and Wales in 1931, the group coming into the inquiry consists of those persons in England and Wales alive on the night of April 26th, 1931, and excludes all those units who do not possess the characteristics of

human beings, it also excludes those human beings who although alive were not in England and Wales on that night. All members of the group possess therefore certain characteristics in like manner.

The information, so described, concerning these units is supplied to us in the shape of filled-in schedules or forms, or as the contents of card index boxes, or as ledgers, or day-books, and so on. Such may be called the raw material of statistics. What are the processes which this raw material goes through before the finished article is produced? These processes are sometimes called "Statistical Methods", and the finished products "Secondary Statistics" as opposed to "Primary Statistics", the result of tabulating the statistical material in its crude form before entering the statistical mill. What is the first process to which these crude data must submit? Obviously to a checking of the accuracy of the information supplied, a careful scrutiny which should establish the existence of any obviously wrong pieces of information, and correction where such seems necessary. After this, when it is felt that the information is trustworthy, the next step is to assemble and condense. No one can appreciate at a glance or even after careful study hold in the mind the information contained in a hundred or a thousand or more schedules, no one, by turning over page after page of a book containing information respecting many units, can get a proper notion of the detail contained there. It is essential that some process of condensation must take place. This process results in Statistical Tables. This tabulation necessarily involves the grouping together of units into classes. The kind of tables produced and the grouping into classes are both determined by the

nature of the information obtained, i.e. by the particular characteristics possessed by the units with which the investigation was concerned. Broadly, we may say that the process of assembly and condensation groups together those units which are alike in respect of certain characteristics. Detail is necessarily lost, the individual unit becomes merged in a group. Few of us can identify ourselves in the Census Tables, for instance, each of us realizes that he or she is merely one of a large number in a particular table. We have to consider what is involved in this process of grouping together like with like, the process called Classification.

Classification is determined by the characteristics possessed by the individual units. These characteristics may be considered as of two kinds, (1) those which may be referred to as descriptive and (2) those which may be referred to as numerical, being susceptible of quantitative appreciation. In the first kind are such characteristics as sex, civil condition, occupation, etc., kind of trade in which employed, in the case of persons; kind of goods carried in the case of ships; type of industry, type of power used, in the case of factories, and so on. In the second kind are such characteristics as age, height, income, rent paid, and so on, in the case of persons; tonnage, number of crew, in the case of ships; value of products, wages bills, rents, and so on, in the case of factories. Some of the characteristics of the first kind may be very easily classified by means of some natural or physical lines of demarkation, and these natural or physical differences have determined the classes into which units possessing this character should be placed. It is easy in these cases to determine whether two units are alike or not in respect

of this character. In this category, for instance, is sex in the case of persons; kind of motive power in the case of ships, whether sail or machines; similarly in the case of vehicles, whether mechanically propelled or horse-drawn; and so on. In these cases the method of classification is obvious, and naturally advantage is taken of this when units are to be grouped together, like with like. But there are many characteristics where this classification is not so easily achieved. For instance there are characteristics which are possessed in varying degrees, which merge into one another when any grading is attempted, such as eye-colour in persons. We may classify eyes as brown and blue, but find cases where the colour is more properly described as grey, we may also decide that we should have a dark and a light brown class, also possibly green, and finally arrange a scale of classes, dark brown, light brown, green, grey, blue. But when we attempt to fit our units into these five classes we may have doubtful cases which we hesitate to describe as dark or light brown, others where it is difficult to decide whether the colour is grey or blue. There are really a large number of eye-colours and if we fix on a definite number of classes, we shall always have border cases where it is not easy to decide to which class such cases belong. So in the case of classifying commodities according to the state of manufacture at which they have arrived when they are being bought and sold. We think vaguely of raw materials and manufactured goods, we recall to mind raw cotton, raw wool, bedsteads, clothing, motor-cars. But when we survey the whole range of goods which enter into trade we realize that there are many grades between raw materials at one end of the scale and manufactured goods

at the other end of the scale. For instance, consider the case of wood. Shall we reserve the words "raw material" to be applied only in the case of trees standing untouched in the forest, or when the tree has been felled, or when the branches have been lopped off, the trunk alone remaining, or when the log has been dragged to the saw mill, or when it is sawn into planks, or when the planks are sawn into standard lengths. In the sense that any one process of manufacture turns raw materials into finished goods, then wood in all these states is at the same time raw material and finished article. The same kind of problem arises when we consider the stages, iron ore, pig iron, steel, rails. Which are raw materials and which are finished goods? After considering problems of this kind, in the end no strict classification into these two groups is really attempted. In the accounts of the Foreign Trade of the United Kingdom, for instance, there are two main groups which are defined as "Raw Materials and Articles Mainly Unmanufactured" and "Articles Wholly or Mainly Manufactured", and for each class of commodity there are further subdivisions into goods of each class which range from raw material to finished product, no definitions involving processes of manufacture being attempted; the distinctions made are those due to description of the goods.

We find, therefore, that in certain cases we can group like with like because there are fundamental distinctions between units, in other cases where classification is difficult we have to be satisfied with grouping together those units which are nearly alike, but we realize that there are probably cases where a unit is placed in one class with others to which he is akin, and does not differ by

much in respect of the particular character from another unit which is placed in a neighbouring class. If we want to condense the original information into manageable proportions we have to be satisfied with this state of affairs. The same sort of difficulties arise when considering classification of units according to those characters of which we can get a quantitative appreciation. In certain cases there is no doubt of the limits of the classes which are used. For instance, if we are classifying households according to the number of persons, these numbers range 1, 2, 3, 4, 5, and upwards. We have no difficulty in saying that one household is like another in this respect, they both contain the same number of persons. Similarly goods trains can be classified according to the number of wagons attached. On the other hand, there are many and various ways in which particular characters may be possessed by individuals. The heights of persons, for instance, are of indefinite variety, so also are ages, or wages. The same is true in the case of tonnage of ships. Where we are dealing with *measuring* as against *counting* we find an infinite number of possibilities, and here again in classification no attempt is made to ensure that like with like go together in one class; this is impossible, all that is done is to put those units together which are nearly alike in respect of the particular character. Thus we group together those whose ages are 20 and more but less than 25, those who are 25 and more but less than 30, and so on. We group together those whose heights are 5 ft. 6 in. and more but less than 5 ft. 9 in., those 5 ft. 9 in. and more but less than 6 ft., and so on.

The question of definition is important in classification. In the first place, the question arises as to the definition

of the units, whether the scope of the inquiry is clearly defined so that all those units, which should form part of the whole group to be considered, are included. Secondly, the characters about which information is obtained are to be clearly defined. Thirdly, the extent to which the units possess these characters must be defined, so that we get those together which are really alike, when we perform the necessary grouping for purposes tabulation. Let us take an illustration of the difficulties which arise from the Census of Population. The problem is to get information relating to the whole population of the country at a particular time, part of this information is to be concerned with some of the environmental conditions. It is decided that the unit, for the most part, shall be the household and that the head of the household shall be responsible for supplying the relevant data. Census forms are therefore to be distributed to all the heads of households in the country. This procedure appears to be simple when we imagine an official progressing slowly along street after street of a town, knocking at the doors and leaving the appropriate form in the correct hands. But the Census official's experience is greater than ours, he has experience of finding a house with two distinct families living there, and he has to have guidance as to whether that is one household or two. If he does not ask for help from those above him, he may decide to leave one form, whereas another official faced with the same problem solves it differently and leaves two forms. This sort of procedure is damaging to the accuracy of the data and must be avoided. The Census official must therefore have help in deciding which groups of people living together are households for Census

purposes and which are not. In other words the Unit of the Inquiry has to be defined efficiently so that no difficulties arise. The Census office gives instructions that the unit household is that group of persons living together in a "Structurally separate dwelling place", a place from which access may be had to the public streets without interfering with or interference from another group of persons. Again, in the Census, information is required as to occupation. In recent years, owing to the prevalence of Unemployment in the country, careful instructions were given so that people could describe themselves properly under this characteristic heading "Occupation". If a man had been a carpenter but had not worked at his trade for a considerable time, but had been employed for brief intervals in this period of unemployment as a market-gardener or a labourer, how should he describe himself on the Census Schedule? The instructions given for the filling up of the Census Form in 1931 were to the effect that if a person were unemployed, but had previously been occupied as a carpenter, he should describe himself as Carpenter (Unemployed), but where a person had no hopes of further employment in his previous occupation, and had engaged himself in a new one, he should describe himself as belonging to the latter category.

Finally, on the Census Form careful instructions are given to assist persons in the filling up of the form so that vague descriptions of occupation may be avoided, in order that, when the subsequent grouping of like with like takes place, there is confidence that persons who are alike in respect of this character have described themselves in the same way.

This emphasis on definition is obviously necessary, otherwise when the process of assembling the data is to be performed, this may be found almost impossible, or the exact meaning of the tables finally obtained may be obscure. Further, the definitions used in an inquiry should accompany the tabular presentation of results in order to render their meaning perfectly clear.

To summarize, the information obtained as the result of an inquiry may be represented thus :—

Inquiry relating to place *X* at time *Y*.

Units	Characters					
	A	B	C	D	E	F
1	A <sub>1</sub>	B <sub>1</sub>	C <sub>1</sub>	D <sub>1</sub>	E <sub>1</sub>	F <sub>1</sub>
2	A <sub>2</sub>	B <sub>2</sub>	C <sub>2</sub>	D <sub>2</sub>	E <sub>2</sub>	F <sub>2</sub>
3	A <sub>3</sub>	B <sub>3</sub>	C <sub>3</sub>	D <sub>3</sub>	E <sub>3</sub>	F <sub>3</sub>
4	A <sub>4</sub>	B <sub>4</sub>	C <sub>4</sub>	D <sub>4</sub>	E <sub>4</sub>	F <sub>4</sub>

*X* and *Y* need definition, the Units need definition ;  
A, B, C, need definition

Unit 1 possesses character *A* to the extent *A*<sub>1</sub>, character *B* to the extent *B*<sub>1</sub>, and so on. Unit 2 possesses character *A* to the extent *A*<sub>2</sub>, and so on.

For example :—

Inquiry relating to London in the month of January, 1933.

Persons	Sex	Age	Wage per week	Civil Condition	Occupation
1	Male	30	53s	Married	Taxi-driver
2	Female	10	—	Single	—
3	Female	28	33s	Married	Waitress
4	Male	55	50s	Married	Railway Porter

London must be defined; there are many "Londons"—the area administered by the London County Council, the Metropolitan Police Area, and so on. Is the month of January, 1933, the calendar month or the four weeks ending 28th January? Under "Wages" are these figures average weekly earnings for the month, are they gross or net earnings, that is, do they include or exclude insurance and mutual benefit contributions, tips, allowances for uniform and travelling, and so on? Is a person to be described as Married who is living apart from his wife?

Note that it is possible that a unit in the group may not possess a particular character at all.

In a subsequent classification of occupations into groups such as Skilled and Unskilled, in which category will such as Taxi-drivers, Waitresses, Railway Porters go?

From these pieces of information in the raw state tables are obtained, where the units are grouped together.

Tables are intended to summarize the information obtained during the course of the inquiry, consequently we never expect these tables to present the whole of the information obtained. Since this is the case, it naturally follows that, owing to the factors of time and labour involved in assembling and condensing the statistical data, a certain amount of choice is necessary as between different matters of interest, so that tabulation is concerned primarily with those particular aspects of an inquiry which will serve most usefully an immediate need. The original material should be preserved, so that it is possible to proceed to further tabulation if the necessity

arises in the future, owing to a transfer of interest to some other problem which previously had not been considered. Since the tabular presentations of statistical data are the first evidences, available to a wider public than the investigators, of the results of an inquiry, the information contained therein should be perfectly clear and precise, that is, explanations should accompany the tables, where necessary, in order that there should be no possibility of ambiguity in interpretation of the meaning of the tables. For example, in the Annual Report of the Secretary for Mines for the year ended 31st December, 1928, Table 25, in the Statistical Appendix relating to Outputs, Costs of Production, Proceeds, and Profits of the Coal Mining Industry during 1928 contains many explanatory notes. In the first place the information relating to South Wales and Monmouthshire is that for the year ending 31st January, 1929. A note explains the sources from which the information is obtained—"The particulars are based partly upon the returns made for the purpose of wage ascertainties for certain districts, and partly upon other returns supplied by individual Colliery Owners." The figures in the table do not refer to the whole Industry but to 96 per cent in the case of the third and fourth quarters of the year and to 97 per cent in the other quarters. There are explanations of what are included in the items "Wages" and "Other Costs of Production".

Tables give information relating to the variety of ways or degrees in which units in a group, which are alike in certain respects, possess one or more characteristics. A simple table contains data respecting one characteristic, the information relating to others not being included. A

more complex table may at the same time contain figures relating to several characteristics. In any of these cases those characteristics possessed in like manner by all the units should be referred to in the heading or description of the table. For instance, in the Table :—

AGES OF MALES MARRIED IN ENGLAND AND WALES IN 1925

Age	Under 21	21-	25-	30-	35-	45-	55 and up- wards	Ages not stated	Total
Number	12,011	97,797	100,550	37,819	26,095	11,203	8,801	1,013	295 689

The 295,689 persons have these characteristics in common : they are of the same sex, they married in the year 1925 in England and Wales ; so they are differentiated from Females, from those who did not marry in that year in England and Wales. On the other hand they were not all of the same age. Their age distribution is indicated in the above table, where there are grouped together as like in respect of this character those whose ages are nearly the same, all those aged 30 and over but less than 35, for example, being placed together in the same group. But these men possess many other characteristics, occupation, height, hair-colour, income, and so on. No information on these points is given in this table. If further information were available (as it is, as to certain characteristics in the registers of marriages) and if a useful purpose were to be served by publishing it, we should subdivide each of these groups into further classes in respect of these other characters. For instance we might have a more complicated table showing Age and Occupation in this way.

Industry in which occupied	Ages	Under 21	21-	25-	30-	35-	45-	55-	Age not stated	Total
	Agriculture									
	Coal Mining									
	Metal Manufacturers									
	Textiles									
	Chemical Manufacturers									
	Others									
	Total									

Each figure in this table would represent a group alike in certain respects, but with individuals differing in other respects and, if further information were available, we could proceed to further tabulation in this way:—  
Distribution of Earnings of Male Coal Miners aged 25 and over but less than 30 who Married in England and Wales in 1925.

## EARNINGS PER WEEK

Un-employed	Under 20s	20s -	25s -	30s -	35s -	40s and over	Total

We should then have a number of tables of this kind, the headings of which would indicate in what respects the units in the group were alike.

In those cases where the number of classes into which units may be grouped in respect of a particular characteristic is large, the amount of information contained in one table is necessarily limited by the size of the paper on which the table is presented, but where the number of classes is small, one table may be used easily to give information respecting a number of characteristics. For example the students attending an educational institution may be divided up into Male and Female, Day Students and Evening Students, First, Second, or Third Year, and they may be classified according to the Course they are pursuing. We may get in a simple table the information relating to these four characteristics in this way :—

STUDENTS ATTENDING THE INSTITUTION IN THE SESSION 1930-1

	DAY			EVENING		
	Men	Women	Total	Men	Women	Total
COURSE A						
1st year	48	35	83	28	13	41
2nd year	52	39	91	36	15	51
3rd year	49	40	89	29	11	40
Total	149	114	263	93	39	132
COURSE B						
1st year	78	73	151	52	24	76
2nd year	70	60	130	48	20	68
3rd year	75	69	144	50	25	75
Total	223	202	425	150	69	219
COURSE C						
1st year	43	68	111	11	4	15
2nd year	28	73	101	15	7	22
3rd year	37	54	91	10	5	15
Total	108	195	303	36	16	52

A table of this kind is presented as a regular feature of the Statistical Abstract issued by the Board of Trade relating to *Shipping engaged in Foreign Trade*. The table gives "Total Net Tonnage of British and Foreign Vessels, distinguishing Sailing and Steam, Entered and Cleared, in the Foreign Trade at Ports in the United Kingdom, with Cargoes and in Ballast and with Cargoes only", for 15 years. The units (ships) possess these four characteristics as to Flag, motive power, Entering or Clearing, and the state of the hold, in varying ways, and the information respecting these matters is easily compressed into a single table in the Abstract.

The construction of a table is determined to some extent therefore by the kind of classification adopted for the purpose of grouping together those units which are alike in respect of a particular characteristic.

In some cases the figures in the tables refer to the number of units, such as in the illustrations above, where the number of males marrying is quoted, or the number of students, but in other cases the figures in the tables refer to some further character, the number of units not being shown at all. For instance, in the Shipping Tables mentioned the numbers in the Table are total net tonnage and not number of vessels. For the purpose of indicating the quantity of shipping engaged in the Foreign Trade the total net tonnage is of more interest than the total number of vessels. Similarly in the Tables referring to the Foreign Trade, the figures are those relating to the total of quantities and values of consignments and not to the number of such. So in the Tables relating to the Coal Mining Industry, the figures generally give Output of Coal in tons and the Number of Men working and not

the number of mines, though where this number is of interest, e.g. in Tables relating to the use of Mechanical Appliances in Coal Mines it is given.

To a considerable extent the detailed form in which the tabulated material is presented depends upon the persons responsible for the construction of the tables, and the factors which influence them are the uses which such tables are to serve. That is to say a table is constructed in the light of some problem or matter of interest. For instance, the following table summarizes the information relating to age at death of persons dying in Great Britain in 1925 :—

AGE DISTRIBUTION OF PERSONS DYING IN GREAT BRITAIN IN 1925

Age in years	Under 5	5-	15-	25-	35-	45-	55-	65-	75 and over	Total
Number	96,329	15,642	22,829	24,410	33,388	53,143	77,800	104,882	111,305	538,548

Such a table may serve satisfactorily to convey roughly the information available as to age at death, but a person interested particularly in Infantile Mortality would find a table of this kind much more useful :—

AGE DISTRIBUTION OF PERSONS DYING IN GREAT BRITAIN IN 1925

Age in years	Under 1	1-	2-	5-	15-	25-	35-	45-	55 and over	Total
Number	62,745	17,279	15,304	15,642	22,829	24,410	33,388	53,143	203,607	538,548

Actually he would probably dispense with the greater part of this table and concentrate on the first group and use a table :—

AGE DISTRIBUTION OF INFANTS UNDER ONE YEAR DYING IN  
GREAT BRITAIN IN 1925

Age in Months	Under 1	1-	3-	6-	9-	Total
Number	26,810	10,446	9,668	8,096	7,726	62,746

We find many different kinds of grouping of ages in tables giving age distributions, these differences being due to the different material tabulated and the variety of uses to which the information is put when tabulated.

This ever-present possibility of variations in the grouping exists wherever we are dealing with a character which can be measured. We can distinguish between two types of tables in this connection. On the one hand there is the kind of table shown in official returns where as much detail is to be presented as is possible, consistent with space in a book and expense, and where it is anticipated that this table will primarily serve the purpose of stating facts, and will be the basis of argument and discussion on the part of persons interested. In such a table we should expect, for instance, that an age distribution would show ages in yearly groups. On the other hand, there are those tables, perhaps derived from official returns, where great detail is not essential, and in fact where too much attention to detail would mar the appearance of the table, and where those interested are expected only to gather the general idea of the variation existing in a group in respect of a particular character. In such cases the individuals put together in a group perhaps range over a wide interval of the variable character. As an extreme instance of this we may consider a simple table of this kind:—

## POPULATION OF ENGLAND AND WALES, 1921

Age in years	Under 15	15-64	65 and over	Total
Number	10,500,455	25,095,139	2,291,105	37,886,699

Here attention is drawn to these particular age-groups representing, more or less accurately, that part of the population of active working life, that part requiring to be supported before they become "units of production", and that part retired from active participation in production. The purpose of the table is to enable comparisons to be drawn between these three figures and the total. This purpose is reasonably well served by such a table, and certainly would not be suited if the table giving the age distribution of the population at every year of age were used. In the latter case we should not be able to see the wood for trees.

When we are dealing with tables conveying information respecting a character, possessed by the units, which is only susceptible of descriptive classification, the same kind of distinctions may be drawn between "official tables" for information only, and those tables which have a place in an argument or discussion, and which are supposed to help to elucidate or clarify a particular point. In the former case the tables again should contain as much detail as possible, and the practice in this respect should be guided by preceding tables showing the results of parallel investigations in previous years, in the first place, because the previous practice was presumably concerned with emphasis on matters of interest and, in the second place, because one of the most necessary

purposes of statistics is to enable comparisons to be made over a period of time. Of course, if circumstances have changed, as they are continually changing, the form of the tables will necessarily be modified so as to be consistent with these changes, but as far as possible the form should be kept as close to the preceding form so that comparisons will not be vitiated, where these are possible. In the latter case the tables should be as simple as possible consistent with the material, and a golden rule is to use two tables instead of one if there is the least fear that one would contain such a mass of material that the essential facts therein would tend to be hidden. Briefly we may say, that after the information is obtained in its crude form, it should be tabulated in as detailed a manner as is considered necessary for the particular data; these tables which result are the primary statistics of the investigation. Subsequently from the primary tables others may be obtained to serve particular purposes in order to emphasize this or that aspect of the problem with which the investigation is concerned. If the inquiry is a simple one, it may well be the case that no further tables are necessary after the preliminary tabulation has been performed, especially if the investigators are quite clear as to what statistical tables are required. But in the case of large-scale investigations, like those instituted by a Government Department, the statistical data might be used for many purposes besides those to which they are immediately put by the particular department, consequently the preliminary tabulations should be as detailed as possible so that they may be of value to later investigators.

## CHAPTER 4

### SECONDARY OR DERIVED STATISTICS

The information obtained by an investigation is now available in the form of tables. These tables are useful to us in that they enable comparisons of various kinds to be made. The purpose of tabulation is partly to present the facts elicited by an inquiry, purely as facts, and partly to present these facts in a manner suitable for the appropriate comparisons. Sometimes a table contains so many figures, and the comparisons which are of interest are those relating one particular figure to many others, that it is not easy at a glance to appreciate all the information contained in the table. Consequently subsidiary tables may be constructed from the original, which serve to emphasize more clearly the matters of interest, and possibly arithmetical processes may be applied to the tables in order to bring these out, or again graphical methods may be employed to attain this end. The application of such arithmetical processes, which may be described as statistical methods, results in what may be called *secondary statistics*.

One of the simplest methods of making comparisons between figures is that of ratios. The ratio method is extensively used in statistical analysis in simple or complicated form and produces secondary statistics which are given different descriptions in different contexts, such as percentages, averages, rates, weighted averages,

TABLE I

NUMBER OF MALES AND FEMALES IN AGE-GROUPS IN 1881, 1901, 1921  
IN ENGLAND AND WALES

Age (years)	1881		1901		1921	
	Male	Female	Male	Female	Male	Female
Under 5	1,757,657	1,763,207	1,855,361	1,861,347	1,681,439	1,640,284
5-10	1,568,579	1,578,817	1,738,993	1,748,298	1,788,560	1,752,368
10-15	1,402,230	1,398,101	1,670,970	1,670,770	1,837,125	1,822,701
15-20	1,268,269	1,278,963	1,607,522	1,638,621	1,727,823	1,775,231
20-25	1,112,354	1,215,872	1,472,644	1,648,278	1,448,385	1,703,067
25-30	981,278	1,066,714	1,328,288	1,496,221	1,339,980	1,820,290
30-35	840,259	905,210	1,157,660	1,273,685	1,281,320	1,519,649
35-40	744,924	796,475	1,034,459	1,110,924	1,273,321	1,471,913
40-45	672,971	726,383	897,484	953,138	1,223,054	1,378,121
45-50	547,508	603,883	759,955	813,233	1,162,158	1,243,988
50-55	485,758	536,317	636,254	692,749	971,021	1,043,130
55-60	381,998	424,468	497,498	535,079	781,608	849,117
60-65	340,555	387,067	410,447	480,226	601,235	680,768
65-70	231,549	290,920	282,403	347,270	449,363	536,699
70-75	158,333	191,622	195,465	250,868	280,491	376,320
75 and over	145,680	190,540	183,204	258,543	250,376	397,856
Total	12,639,902	13,334,537	15,728,613	16,799,230	18,075,239	19,811,460

index-numbers. The purpose in every case is the same, to reduce the comparable figures to such proportions that the appropriate comparisons are more easily made. Suppose, for instance, that we were concerned with such a table as that on page 43 which gives information of the age distribution of the male and female population of England and Wales at three census dates 1881, 1901, 1921.

This table serves the purpose of emphasizing the increasing numbers in the population in 40 years, it emphasizes the numerical superiority of females over males, and the larger number of young persons compared with old persons. But much remains hidden until drawn out by means of secondary statistics. We cannot get a proper appreciation of the fact that the constitution of the population is in a state of flux, until we have a table of derived statistics which results from making comparisons. It is appropriate to compare any one figure in this table with several others, for instance, we can compare the number of Male Children under 5 years in 1881 with the total Males in 1881, with the number of Females under 5 in that year, and with the number of Males in that age-group in 1901 and 1921. These comparisons will be related to three different matters of interest to those concerned with the size and distribution of the population. We therefore construct new tables giving ratios in the form of percentages, as on pages 45 and 46.

Table I (a) emphasizes the age constitution of the population by concentrating on the proportions in different age-groups. For instance we appreciate, by glancing at this table, the large part of the population under 15 years of age. Further, this table enables comparisons in this regard to be made between the male and female population,

TABLE I (a)

PROPORTIONATE AGE DISTRIBUTION OF POPULATION IN ENGLAND  
AND WALES IN 1881, 1901, 1921

Age (years)	1881		1901		1921	
	Male	Female	Male	Female	Male	Female
Under 5	13.9	13.2	11.8	11.1	9.3	8.3
5-	12.4	11.8	11.1	10.4	9.8	8.8
10-	11.1	10.5	10.6	9.9	10.2	9.2
15-	10.0	9.6	10.2	9.8	9.6	9.0
20-	8.8	9.1	9.4	9.8	8.0	8.6
25-	7.8	8.0	8.4	8.9	7.4	8.2
30-	6.6	6.8	7.4	7.6	7.1	7.7
35-	5.9	6.0	8.8	6.6	7.1	7.4
40-	5.3	5.4	5.7	5.7	6.6	7.0
45-	4.3	4.5	4.8	4.8	6.4	6.8
50-	3.8	4.0	4.0	4.1	5.4	5.3
55-	3.0	3.2	3.2	3.3	4.3	4.3
60-	2.7	2.9	2.6	2.9	3.3	3.4
65-	1.6	2.0	1.6	2.1	2.5	2.7
70-	1.3	1.4	1.2	1.5	1.6	1.9
75 and over	1.2	1.4	1.2	1.5	1.4	2.0
Total	100	100	100	100	100	100

TABLE I (b)

NUMBER OF FEMALES PER 100 MALES IN EACH  
AGE-GROUP IN ENGLAND AND WALES

Age (years)	1881	1901	1921
Under 5	100.3	100.3	97.5
5-	100.7	100.5	99.2
10-	99.7	100.0	99.2
15-	100.9	101.9	102.7
20-	109.3	111.9	117.6
25-	108.7	112.6	120.9
30-	107.7	110.0	118.6
35-	106.9	107.4	115.6
40-	107.9	106.2	112.7
45-	110.3	107.0	107.0
50-	110.4	111.4	107.4
55-	111.1	111.6	108.6
60-	113.7	117.0	110.7
65-	117.0	123.0	119.4
70-	121.1	128.3	134.2
75 and over	130.7	141.1	158.9
Total	105.5	106.8	109.6

TABLE I (c)

NUMBER OF MALES AND FEMALES IN EACH AGE-GROUP IN 1901, 1921  
AS PERCENTAGES OF THOSE IN 1881 IN ENGLAND AND WALES

Age (years)	1881		1901		1921	
	Male	Female	Male	Female	Male	Female
Under 5	100	100	105.6	105.6	95.7	93.0
5-	100	100	110.9	110.7	112.6	111.0
10-	100	100	119.2	119.5	131.0	130.4
15-	100	100	126.7	128.1	136.2	138.8
20-	100	100	132.4	135.6	130.2	140.1
25-	100	100	135.4	140.3	136.6	151.9
30-	100	100	137.8	140.7	152.5	167.9
35-	100	100	138.9	139.5	171.0	184.8
40-	100	100	133.4	131.2	181.8	189.7
45-	100	100	138.8	134.7	212.3	206.0
50-	100	100	131.0	129.2	199.9	194.5
55-	100	100	130.2	130.8	204.6	200.0
60-	100	100	120.5	124.1	176.5	175.9
65-	100	100	122.0	128.2	194.1	198.1
70-	100	100	123.5	130.9	177.2	196.4
75 and over	100	100	125.8	135.7	171.9	208.8
Total	100	100	124.5	126.0	143.0	148.6

and different years. As between male and female, we note, for instance, that the figures for females are less than those for males in each age-group under 20 years, and more in the age-groups over 20 years. This is true for the three years 1881, 1901, 1921. Except in the latter year in respect of age-groups 45-, 50-, the age distribution of the female part of the population is definitely different from that of the male part of the population. This fact certainly cannot be appreciated by glancing at Table I. Tables I (b) and I (c) further bring out the contrast between male and female. Women appear to live longer than do men, consequently the proportions of females to males in the older age-groups is higher than in the younger age-groups. This is brought out in Table I (b). Further,

comparisons are possible between 1881, 1901, 1921 and any changes with time in this respect may be therefore observed. The effect of the War is apparent in the proportions of females to males in the age-groups 20-45

TABLE II

NUMBERS ENGAGED IN CERTAIN INDUSTRIES IN GREAT BRITAIN  
1911 AND 1921. OCCUPIED MALES AGED 10 YEARS AND OVER

Industries	1911		1921	
	Number	Per-centage	Number	Per-centage
Fishing . . . . .	63,000	0.5	63,000	0.5
Agriculture . . . . .	1,301,000	10.1	1,198,000	8.8
Coal and Shale Mining	1,122,000	8.7	1,294,000	9.5
Manufacture of Bricks, Cement, Pottery, and Glass . . . . .	162,000	1.2	161,000	1.2
Manufacture of Chemicals, Explosives, Paints, Oils, Rubber, etc . . . . .	145,000	1.1	195,000	1.4
Manufacture of Metals, Machines, Implements, and Conveyances . . . . .	1,670,000	12.9	2,251,000	16.5
Manufacture of Textiles . . . . .	585,000	4.5	540,000	4.0
Manufacture of Cottons . . . . .	257,000	2.0	234,000	1.7
Manufacture of Wool and Worsted . . . . .	118,000	0.9	115,000	0.8
Manufacture of Silk . . . . .	11,000	0.1	14,000	0.1
Manufacture of Flax, Hemp, Jute, Rope, Canvas, and Canvas goods . . . . .	40,000	0.3	33,000	0.2
Manufacture of Dyeing, Bleaching, Printing, and Finishing . . . . .	91,000	0.7	89,000	0.6
Total Occupied . . . . .	12,930,000	100	13,856,000	100

in the 1921 figures of the table. Table I (c) concentrates on the growth in population in the four decades and we may note particularly, for example, that whereas the female population as a whole has increased in 40 years

by about 50 per cent, the older part of it, including those over 45, has increased by about 100 per cent. It is not the object, here, to discuss these figures at great length, the emphasis is laid on the fact that, in order to appreciate the information contained in a table like I above, it is necessary that subsidiary tables like I (a), I (b), and I (c) should be prepared. This method of procedure is general. If we are concerned with a table which consists merely of two or three figures, it is possible that the relationships between them may be sufficiently evident without further calculations, but where a table involves a mass of figures, it is practically always necessary to reduce these, somehow, to ratio form. It is therefore usual in presenting tables of this kind to include in the tabulated information, the results of calculations of this nature, in order that those wishing to obtain a proper appreciation of the figures may do so without somewhat burdensome calculations. As an illustration the table on p. 47 may be cited.

The effect of such calculations as those which have been illustrated is to enable comparisons to be made by reducing the figures we are concerned with to what may be considered as more manageable proportions, or to reduce a series of figures to a common denominator

The same kind of notion is in mind when we calculate those ratios which are known as averages. Here we are generally concerned with the ratio of numbers which are expressed in different units, and we find how much of one quantity would accrue to each individual unit of the other quantity in the whole group, if the distribution were equal as between individuals. For instance, in the following table figures are given respecting importation

of certain articles for home consumption in the United Kingdom :—

TABLE III

HOME CONSUMPTION OF IMPORTED ARTICLES INTO UNITED KINGDOM, 1911, 1921

Articles	1911		1921	
	Quantities	Quantities per head of population	Quantities	Quantities per head of population
Butter	cwt 4,167,140	lb 10.31	cwt 3,329,418	lb 7.91
Wheat, Grain, and Flour in equivalent of Grain	cwt 111,497,952	lb 275.86	cwt 99,184,732	lb 235.74
Eggs	thou 2,265,804	No 50.06	thou 1,263,680	No. 26.83
Beef, Fresh and Refrigerated	cwt. 7,315,333	lb 18.10	cwt 10,972,014	lb. 26.08
Mutton and Lamb, Fresh and Refrigerated	cwt 5,322,159	lb. 13.17	cwt 6,811,617	lb. 16.19
Bacon and Hams	cwt 5,681,307	lb 14.06	cwt 6,255,717	lb. 14.87
Estimated Population	45,268,000		47,123,000	

In this table comparisons are made between the imported quantities and the total population, the result of such comparisons being given as "average per head of population". The effect of the changing population is thus eliminated and comparisons between the average in 1911 and 1921 enable effective deductions to be drawn as to the changing volume of imports over this period of years. The average presents a crude picture of the amount imported for consumption, by indicating what each individual person in the whole of the United Kingdom would receive if these imports were distributed evenly. Similarly, we speak of the average wage of a group of men, obtained by dividing the total wages received by the total number of men who receive these wages, and this average is a figure representing what each would have if the amount was divided equally. So also, we work

out the average rent of a group of houses, dividing the total of rents by the number of houses. An average is, then, obtained as a ratio in the form Numerator  $\div$  Denominator, where the Numerator represents the total extent to which a particular characteristic is possessed by the whole of a group, and the Denominator represents the total number in the group, or it may be the total extent to which another characteristic is possessed by the whole of the group. For instance, in railway statistics the average wagon load is calculated for a particular area of railway operation over a certain period of time. This is obtained by dividing the "ton miles", which is got by summing the results of multiplying each item of freight by the distance hauled, by the "loaded wagon miles" obtained by totalling the distance moved by each wagon (loaded). So, in the Coal Mining Industry, the average output of coal (in tons) per manshift worked is calculated for particular areas over periods of time, by dividing the output by the total number of manshifts worked. In both these cases neither the numerator nor denominator are totals of the original units, which are freight trains in the first example and coal mines in the second, but are totals of characters possessed by these units.

Sometimes the ratios obtained when making comparisons are called "rates" and may be expressed as "rate per cent" or "rate per mille" or "rate per thousand". Whichever of these is used is merely a matter originally of convenience, and latterly of precedent. We use for instance birth rates and death rates which are obtained by relating as numerator to denominator the births or deaths in a given period (usually a year) to the total

population, and expressing the result as a "rate per thousand". These derived statistics serve to indicate the natural increase and decrease of the population, and enable comparisons to be made from one time to another.

Now that we have arrived at the stage of performing arithmetical calculations on our statistics some discussion is appropriate on the subject of accuracy. It is unlikely, when we are finding the ratio of any one number to another, that we shall get the quotient, as we do in easy arithmetical exercises, exactly without remainder, and the question naturally arises as to how many decimal places the answer is required; in other words, what degree of accuracy do we want in the result. This question can never be answered in set terms, the answer depends on the data which are to be submitted to this process and on the particular problem. For instance, the figures in Table I (a) were given to one decimal place, and by doing so the table quite satisfactorily serves its purpose; no advantage would accrue if the percentages had been worked out to two places of decimals. But if this table were to serve another purpose than that in this context, it might be better if greater accuracy were required; if, for instance, the figures in that table were themselves to submit to arithmetical processes of division, then those figures should be calculated to two or three decimal places. In some cases, what matters is the extent to which figures given as a result of calculations can be appreciated by others for whom the work is intended. For instance, a number of things may be divided up into three groups in this way:—

Group	A	B	C	Total
Percentage	28	54	18	100

This table adequately represents the grouping. If these percentages were shown in this form :—

Group	A	B	C	Total
Percentage	28.32	53.87	17.81	100

no advantage is gained. In fact there is a loss of efficiency on the part of the table as a vehicle of expression of a result, because persons reading the table have to substitute 28 for 28.32, 54 for 53.87 and 18 for 17.81 when trying to understand it, and there is no doubt that this process of approximation is gone through. On the other hand, the table might be performing two functions at the same time, bringing out the different ways in which the total is distributed through these three groups, and at the same time emphasizing the fact that as far as two of the groups are concerned the distribution is nearly the same. For instance, a table of this kind :—

Group	A	B	C	Total
Percentage	29.17	28.64	42.19	100

might be better from this point of view than :—

Group	A	B	C	Total
Percentage	29	29	42	100

On occasions, it is more important to consider the number of significant figures to which a result is given rather than the number of decimal places involved. After all, the number of decimal places can readily be altered by changing the wording of the phrase of which the ratio we are considering is a part. Thus instead of speaking of a birth rate of 18.3 per thousand of the population, we can say a birth-rate of 183 per ten thousand, and the decimal place disappears. The total value of Imports into the United Kingdom in 1925 is given as £1,320,715,190 to the nearest £. This figure might be given as £1,321,000,000 to the nearest million, or to four significant figures. The birth-rate in England and Wales in 1925 is 18.3 per thousand (to three significant figures). The percentage figures given in Tables I (b) and I (c) are given to four significant figures. The accuracy of a result is indicated roughly by the number of significant figures involved. The figure 18.3 quoted above as the birth rate in 1925 may be any number from 18.250 to 18.349 inclusive we could only find out where it actually is within this range by recalculating this rate from the original figures of births and population. Thus roughly, at the outside, using 18.3 instead of a more correct figure may involve an error of .05 in excess or .05 in defect, i.e. .05 in relation to 18.3, or about 3 in a thousand. In statistical work the number of significant figures in a result is indicated by the number of figures quoted, thus in Table I (c) the figure 131.0 occurs, the ratio has been calculated to four significant figures, the last being 0. If the ratio had been calculated to three significant figures only it would have been expressed as 131. Similarly, in large numbers such as those in Table II giving numbers in

different industries, the figures there are approximate only, to the nearest thousand, and this approximation is indicated by the presence in each of three zeros.

But sometimes the degree of accuracy in the result of calculating ratios is *not merely* determined by the context to which this result relates, but by the degree of approximation involved in the figures which form the numerator and denominator of the ratio. If these two figures are only accurate to a certain degree of approximation, the resulting ratio is certainly only approximately correct within certain limits. Thus, if numerator and denominator are both 2 (one significant figure), the ratio obtained (1) is not correct to this figure, because the numerator and denominator may be any numbers between, roughly, 1.5 and 2.5, unless we have other information giving these figures to two significant figures, and the ratio will be somewhere between  $\frac{1.5}{2.5}$  and  $\frac{2.5}{1.5}$ , i.e. between

0.6 and 1.7. Thus to one significant figure the ratio is either 1 or 2. In the same way if we are obtaining the ratio of  $\frac{3.4}{2.5}$  where numerator and denominator are given

to two significant figures, the result 1.36 must be considered together with the outside limits which this ratio might achieve, if 3.4 stands for any number between 3.35 and 3.45, and 2.5 stands for any number between 2.45 and

2.55. These limits are  $\frac{3.35}{2.55}$  and  $\frac{3.45}{2.45}$  i.e. 1.314 and 1.408,

1.36 being nearly half way between them. To say that the result is 1.36 is certainly wrong because this implies precision in the result of a degree lacking in the original figures, the result might be 1.38 for instance. To say

that the result is 1.4 is wrong, because the result might be 1.33 which is expressed as 1.3 to two significant figures. All we can do is to say that the result is 1, or we can say  $1.36 \pm .05$ , in this way indicating the limits to which the ratio might reach when the result is indicated to two places of decimals, or we might say 1.36 with a possible error either way of  $3\frac{1}{2}$  per cent.

Two important matters emerge from this discussion of the accuracy of ratios. In the first place, if we are content with percentages, averages, rates, and such like containing only a small number of significant figures, and in many pieces of statistical work these are quite satisfactory for the purpose, then we do not need to insist on absolute accuracy in the numerator and denominator. For instance, in a particular quarter in the Coal Mines of Great Britain 61,833,281 tons of coal are raised by men working 58,218,785 man-shifts in that period and the average output per man-shift is given as 21.24 cwts. We can equally well obtain this figure from the information that 61,833,000 tons were raised, the number of man-shifts being 58,219,000. Now, the number of decimal places in this result (21.24) is quite enough for practical purposes, therefore from the practical statistical point of view we are content to know the output and the number of man-shifts to the nearest thousand. This is important because we may well imagine that the original figures of this kind may be subject to slight errors in counting, perhaps a ton of coal has escaped attention, and from the practical point of view when making comparisons between these large numbers we do not mind if this kind of error has entered so long as it does not amount to much. Consequently, whereas in accounting every unit figure has

a place because items and totals must check, in statistics, where the figures are used for purposes of comparison and where the comparisons are effected by making ratios, since absolute accuracy is *not* essential in these, round numbers or approximations are good enough in the original figures. This does not mean that we should not try for accuracy in the original data ; not at all accuracy is necessary, but we need not quote these numbers to all the significant figures which are involved in the originals, the arithmetical labour is thereby reduced.

In the second place, on many occasions we find that difficulties are involved in obtaining absolutely accurate figures in the first instance. For example, errors are likely to enter into figures of the Import and Export trade ; we cannot be absolutely certain that *all* the persons in the country have been enumerated in the Census, or that some person has not been counted twice ; we cannot absolutely rely upon every householder giving correctly the number of living rooms in his house, thus the total number of rooms may be wrong ; and similarly in any investigation difficulties arise which make certain of the figures in a published table suspect. Moreover there are cases where the information used is not obtained directly as the result of an investigation *ad hoc*, but indirectly through some other means. For instance, the cost and trouble of a Census of Population prohibits its being taken at very frequent intervals, but for many reasons it is useful to know what the population is subsequent to the last Census, and it is interesting to forecast the future population of a country. Consequently, estimates of the population are made each year, and these, *being estimates*, do not pretend to accuracy, they are

given to the nearest thousand. Similarly estimates are made each year by the *Ministry of Labour* of the numbers of workers insured under the Unemployment Insurances Acts, these estimates being used in the Calculations of the Unemployment Percentages issued monthly. These estimates are to the nearest ten, they do not pretend to accuracy. Similarly, the Board of Trade publishes annual estimates of the imports and exports of services (as contrasted with goods) which enter into the Balance of Trade. Now in all estimates of this kind it is no use pretending to be accurate, consequently any ratios calculated with such a figure as a member cannot pretend to accuracy beyond a certain number of significant figures, limited by the number of significant figures in the estimate. The same applies, if it is felt that although accuracy has been urged in an investigation, this has not resulted. It is therefore fortunate that, in practice, we are content with our resultant ratios to a small number of significant figures.

Round numbers or approximations are therefore used extensively in statistical work, especially when *large* numbers are involved, and these limit the amount of arithmetic involved in calculations and yet give us results which are exact enough for practical purposes. But, when round numbers are used, care must be taken that the resulting ratios are not worked out to a greater degree of apparent accuracy than the original figures warrant.

NOTE :—When logarithms are used in the calculation of ratios, 4-figure logarithms give results correct to 3 significant figures, 5-figure logarithms give results correct to 4 significant figures.

## APPENDIX

## TESTS OF RANDOM SAMPLING

In Chapter 2 we mentioned the possibility of testing whether a sample was representative of the group from which the sample was obtained.

Each unit in the whole group possesses a number of characteristics (say) A, B, C, . . . Information respecting a number of these (say) A, C, P, Q, R, is obtained from those units coming within the scope of the sample. Now some information of this kind is probably available concerning the whole group from which the sample is taken. For instance we may know all about characters A and C. If this is the case, we can compare the sample with the whole group as far as these two are concerned. This would be in the nature of a comparison of averages or ratios. We would compare (say) the proportion in the group possessing character A with that in the sample with this character. Or we might find the average amount of C in the whole group and compare it with the average amount of C in the sample. If the sample is representative then these proportions or averages should be the same or nearly the same within reasonable limits which can be calculated. (The theoretical considerations involved in the determination of these limits are too complicated to be dealt with here.)

This testing of the sample should always be possible, because, if we know sufficient to be able to identify the units in the whole group, we are likely to have some knowledge of certain characters possessed by this group. We can always arrange that, when sampling, we should obtain the information respecting *these* characters in

addition to the other information we want from the units in the sample.

For instance, suppose we are taking a sample of households in a certain town. In the sample there will be a certain proportion of school children. Now the Local Authority possesses information respecting the total number of these in the town, and the proportion of school children to total population can be calculated. The proportion in the sample should agree within certain limits with this figure. Or suppose we are taking a sample of the Insured Workers, there should be in the sample the same or nearly the same proportions of persons in different Industries, Coal Mining, Building, etc., as there are in the whole group of Insured Workers.

(See *Livelihood and Poverty* and *Journal of the Royal Statistical Society*, 1924, p. 544 ; 1928, p. 519.)

## CHAPTER 5

### COMPARISON OF AVERAGES

The ratios which are calculated with the idea of effectively comparing one figure with another, and which are called percentages, averages, and so on are themselves subject to comparisons later. In all the illustrations given, such comparisons of secondary or derived statistics are possible. For instance we compare death rates in one town with those in another, or we compare death rates over a period of time. We may compare the import of wheat per head of population with the import of beef, or the import of wheat per head over a period of time. It is important to realize that we are concerned not only with the secondary statistics, but with the primary statistics also. This fact is often forgotten. We must remember that  $(\text{Ratio})_1$  compared with  $(\text{Ratio})_2$  means comparing  $\frac{(\text{Numerator})_1}{(\text{Denominator})_1}$  with  $\frac{(\text{Numerator})_2}{(\text{Denominator})_2}$ . By concentrating on the ratios we are apt to forget that they were derived from numerators and denominators, and that these refer to groups of units. Now ratios may change for a variety of reasons and one of them is, that the groups themselves may change in constitution. For instance, we may take an illustration from the Coal Industry. One of the "efficiency indicators" in this Industry is "Output per Unit of Labour" or average output per manshift worked. Now this average may increase for the whole of the industry from one year to another. This increase might be due to more efficient use of man-power, say,

the increased use of machines for coal-getting. But it may also be due to a change in the number of mines working. Suppose, at the later date the relatively less efficient units of production are no longer operating, but that in any mine which is in operation at both periods the same output per manshift is obtained, then the manshifts worked which form (Denominator), will be manshifts in the relatively more efficient mines, and the second ratio will be greater than the first. Thus, the change in the ratio may merely be indicative of a difference in the constitution of the group as between the two periods under review. As another illustration, consider a concern in the catering trade which sells meals, chocolates, tobacco, wines. A useful secondary statistic to calculate from time to time is the ratio of money received by sale of its goods, to the money expended on purchases of the raw materials at wholesale prices. Such a ratio must be high to allow for salaries, wages, heating, lighting, depreciation, and so on. Now the margin of profit which is possible will vary as between different classes of goods sold, the highest will be on those articles of food which submit to cooking operations and other services, and the lowest on (say) tobacco, the retail price of which is determined by outside competition and agreements with wholesalers from one time to another. Then the ratios worked out in the manner described might show changes which are due to changes in the volume of consumption of the different classes of goods, and may not be due to any change in efficiency of management, whereas it may be the original purpose of these ratios to reflect such changes as the latter. Let us illustrate this with a numerical example. Suppose at one time we have the following figures :—

	Food	Chocolate	Tobacco	Wine	Total	Per cent
Cost	£1,000	£100	£200	£700	£2,000	100
Sale	£3,000	£125	£230	£1,260	£4,615	232

At a later time the following figures are obtained .—

	Food	Chocolate	Tobacco	Wine	Total	Per cent
Cost	£1,000	£100	£300	£600	£2,000	100
Sale	£3,000	£125	£345	£1,080	£4,550	227

The change from 232 to 227 is due entirely to the figures making the totals, and the margins of profit obtained in the various trading departments are unchanged. Similarly, difficulties arise when comparing death rates of one community with those of another. It is well known that old people are more likely to die in a given year than younger people, that very young infants are more likely to die than children who have survived the first year. The death rate of any community is partly determined by the age constitution of the group, and a difference between two death rates may be due in part or wholly to the different proportions in different age-groups in the communities, the death rates of which are under discussion. Similarly the average consumption of tobacco per head of population may change from one period to another partly because the proportion of the population who are smokers may have changed considerably.

The point at issue is one of interpretation of the calculated ratios. The fact that a ratio has changed may be known. What does this mean? Is this change

due to a corresponding change in the individual unit's possession of a particular character, or is the change due entirely or partly to some change in the constitution of the group formed by the individual units? The tendency, in general, is to attribute the change to the first cause, and it is only on further examination of the statistics that it may be found that the second cause is also contributory. This difficulty is met, in practice, by splitting up the whole group into constituent parts when this is possible, and where it is felt to be desirable. For each part, in which the units may be considered alike in respect of the character or characters under discussion, separate ratios are calculated and comparisons are made between these ratios, subgroup by subgroup. Thus, if we wished to compare death-rates in one community with those in another, with the idea of finding out the effect of fundamental racial or environmental conditions on this question of the likelihood of dying within a year, we should compare the death-rates of those in specified age-groups, and so evade the complication due to the fact that there may be in the different populations different proportions at various ages. Further we might consider the question of regrouping into occupations, because different occupations may have different death-rates, and a different occupation-constitution of the population may be a contributory cause to a change in the death rate. Similar remarks apply to marriage-rates. The fact of the marriage-rate changing may be due to change in custom, or it may be due entirely to a change in the age-constitution of that part of the population eligible for matrimony. Certainly before we can assert that such a change is due to the first cause, the ground has to be

examined to find out whether the second cause is not the main reason for the change observed

We, therefore, split up the original group into subsidiary groups which contain units homogeneous or alike in respect of some character possessed by the individuals in different degrees, this character being suspected or known to be connected in some way with other characters, which are the subject of discussion by the consideration of changes in ratios. Thus, if we really want to find out about changes in custom in a community in respect of consumption of tobacco, for instance, we should consider the problem from two angles, first, dealing with the changes in the proportion of smokers in the population over a series of years, secondly, dealing with the average annual consumption per head of that part of the population consisting of smokers, instead of considering the crude figures obtained as average consumption per head of the population. Unfortunately, of course, the number of smokers in the population is not known. Similarly, if we wish to analyse changes taking place over a period of time in the birth-rate, we relate the number of legitimate births year by year to the number of married women (ages 15-45), and get the average number of births per 100 married women of childbearing age. So with marriages, we get a better figure than the crude marriage rate, if we relate the number of men (say) marrying in a year to the number of bachelors and widowers of marriageable age available in that year. We try to get the numerator and denominator of our ratio so related, that the result will give correct information about the question under consideration, and will not be influenced by extraneous factors. As a further illustration the

table below is given. The Monthly Railway Statistics issued by the Ministry of Transport include tables from which these figures were extracted :—

## G B RAILWAYS

Average Wagon Load (tons). July, 1929 and 1930

Railway Company and Area.	Class of Freight							
	I Merchandise and Live-stock		II Minerals and Merchandise		III Coal, Coke, and Patent Fuel		All Freight	
	1929	1930	1929	1930	1929	1930	1929	1930
Great Western	2.92	2.91	9.18	9.10	9.97	10.04	5.84	5.66
Western	2.56	2.59	8.47	8.93	9.47	9.62	4.58	4.62
Midland	2.58	2.64	9.02	9.46	9.21	9.40	4.82	4.83
South Wales	4.48	4.25	10.19	8.78	10.45	10.45	8.40	7.98
London & North Eastern	2.83	2.75	9.51	9.17	10.01	9.96	5.81	5.68
Southern (Eastern)	2.60	2.65	8.79	9.01	9.18	9.12	5.10	5.20
Southern (Western)	2.74	2.68	9.35	9.42	9.44	9.53	5.98	5.97
North Eastern	2.98	2.88	10.49	10.81	12.48	12.47	5.23	5.84
Southern Scottish	3.03	2.97	8.49	8.53	8.90	8.89	5.22	5.03
Northern Scottish	2.68	2.55	6.92	7.37	8.68	8.75	4.14	4.14
London, Mid & Scottish	2.92	2.80	8.87	9.04	8.77	8.62	5.24	5.15
Western	2.94	2.96	8.82	9.04	8.92	9.00	5.02	4.96
Central	2.63	2.65	8.09	8.39	9.19	9.32	5.81	5.44
Midland	2.53	2.58	9.18	9.27	8.61	8.64	5.50	5.47
Northern (South)	3.42	3.28	8.62	8.69	8.79	8.71	5.26	4.87
Northern (North)	2.98	2.21	7.84	7.87	8.76	8.72	4.04	3.81
Southern	2.69	2.73	8.43	8.39	9.47	9.47	4.92	4.88
Chesbire Lines Committee	3.06	2.96	9.29	9.40	9.17	9.26	5.39	5.31
Metropolitan	2.77	2.79	8.79	9.01	9.04	9.30	5.23	5.41
Midland & G.N. Joint	2.09	2.30	8.86	9.23	8.83	9.02	4.46	4.76
Great Britain	2.83	2.80	9.10	9.15	9.49	9.49	5.51	5.41

The average figure worked out for the whole of Great Britain has changed from 5.51 tons (July, 1929) to 5.41 tons (July, 1930). But it is reasonable to suppose that this figure will vary from district to district owing to the characteristic distribution of industry throughout the country, and in fact we find the figure highest in South Wales, and lowest in Northern Scotland. Moreover it would be anticipated that this figure would be influenced by changes in volume of different kinds of freight traffic.

We find in fact that, when freight traffic is divided into the three main classes into which the railway companies group their traffic, the figures are fundamentally different from one another. The load in the case of minerals and coal is much greater than in the case of general merchandise. It is therefore preferable, if a comparison is to be made between operations in these two periods, that the average wagon load should be separately calculated, as in the table, for different areas and different classes of merchandise. A comparison of these figures enables us to decide whether any change has taken place in loading of wagons, which would not be obtained from the crude figures obtained with respect to the whole operations. Let us look at the figures relating to the operations on the London Midland and Scottish Railway (Western Section) reproduced here :—

I		II		III		ALL FREIGHT	
1929	1930	1929	1930	1929	1930	1929	1930
2.94	2.95	8.82	9.04	8.92	9.00	5.02	4.96

The average figure for all freight traffic has declined, we might conclude that the loading of wagons has not been so efficiently performed in July, 1930, as in July, 1929, but the individual averages for the different kinds of freight all show higher results. The average for all freight is less than before purely and simply because there has been a change in the relative volume of goods of the different classes carried.

In these Railway Statistics we subdivide the whole group into a number of sections and obtain ratios for

each, implementing the principle laid down in this discussion of obtaining a ratio from a numerator and denominator which relate to a homogeneous group. In cases where this cannot be done, for instance in the case of the average consumption of tobacco per head of population, where we do not know the proportion of the population which consists of smokers, we have to be content with the crude ratio, which is open to the objection that the apparent reason why the ratio changes may not be the real reason, so that any conclusions drawn must be merely suggestive and not definite.

In many cases, where the sort of difficulty discussed in the preceding pages arises, it is possible to overcome it by a simple device which renders possible a comparison between two groups by means of a single ratio, instead of by a number of ratios. Suppose we consider for illustrative purposes those figures in the table above already discussed relating to the London Midland and Scottish Railway (Western Section) in July, 1929, and July, 1930. If we examine the sources from which these figures were obtained we have the following information :—

CLASS OF FREIGHT

JULY, 1929, 1930	I		II		III		ALL FREIGHT	
	1929	1930	1929	1930	1929	1930	1929	1930
Net Ton Miles (mn) . . . . .	82	80	67	59	66	62	215	201
Loaded Wagon Miles (mn) . . . . .	27·8	27·1	7·6	6·5	7·4	6·9	42·8	40·5
Average Wagon Load (tons)	2·94	2·95	8·82	9·04	8·92	9·00	5·02	4·96

(Average Wagon Load = Net Ton Miles ÷ Loaded Wagon Miles)

NOTE.—Only 2 or 3 significant figures are given in the above table.

This table shows why the average wagon load for all freight is less in July, 1930, than in July, 1929, although the averages for Classes I, II, III have increased in this year's interval. There has been a heavy reduction relatively in Class II and Class III compared with Class I in wagon miles and ton miles, consequently Class I becomes relatively more important in the total at the later date, and since the average wagon load in this class is small compared with the other average figures, the general average is depressed. But suppose we relate the average wagon loads of these classes to some constant standard distribution of ton miles or wagon miles between these three classes, we can, by referring to this hypothetical distribution, find averages which will only reflect in their changes, those changes due to loads. Let us take the distribution of loaded wagon miles for the whole year 1929 as standard; in this year the distribution between the three classes was:—

YEAR 1929 LOADED WAGON MILES (MN)			
Class of Freight			
I	II	III	ALL FREIGHT
325	82	97	504

Consider the following Scheme:—

Class	(1) Loaded Wagon Miles, 1929 (mn)	(2) Average Wagon Load, July, 1929 (tons)	(3) Tons Miles (1) × (2) (mn)	(4) Average Wagon Load, July, 1930 (tons)	(5) Ton Miles (1) × (4) (mn)
I	325	2.94	955.4	2.95	958.8
II	82	8.82	723.2	9.04	741.3
III	97	8.92	865.2	9.00	873.0
ALL	504		2543.8		2573.1
Standardized		JULY, 1929		JULY, 1930	
Average Wagon Load		2543.8		2573.1	
(All Freight)		504		504	
		= 5.05 tons		= 5.11 tons	

We calculate what the ton miles would be for each class of freight, given the average wagon loads of July, 1929, and the loaded wagon miles of the standard year 1929. From that total we get the average wagon load, 5.05 tons ; which we may call the standardized average wagon load for July, 1929. Similarly we get the standardized average wagon load for July, 1930. We can make these comparisons :—

AVERAGE WAGON LOAD (TONS) JULY, 1929 AND 1930

	July, 1929	July, 1930
Crude . . . . .	5.02	4.96
Standardized . . . . .	5.05	5.11

When comparing 5.02 tons with 4.96 tons, the comparison is vitiated because we have to take into account the change in volume of traffic between the classes of freight, but 5.05 tons compared with 5.11 tons gives us a direct comparison, independent of this change. We are enabled, in this way by standardizing the distribution of traffic, to get averages for all freight which will reflect changes due to one cause alone. Of course, the actual standardized averages calculated will depend partly on the period chosen as standard, but the comparison will not be vitiated on this account, and the comparison is the important point to be considered. For instance we might take the distribution in loaded wagon miles in the year 1930 as standard :—

YEAR 1930 LOADED WAGON MILES (MN)  
Class of Freight

I	II	III	ALL FREIGHT
315	75	92	482

Consider the following Scheme:—

Class	(1) Loaded Wagon Miles 1930 (mn)	(2) Average Wagon Load, July, 1929 (tons)	(3) Ton Miles (1) × (2) (mn)	(4) Average Wagon Load, July, 1930 (tons)	(5) Ton Miles (1) × (4) (mn)
I	315	2.94	926.0	2.95	929.2
II	75	8.82	661.5	9.04	679.0
III	92	8.92	820.8	9.00	828.0
ALL	482		2408.3		2436.2
<div> <div>Standardized Average Wagon Load (All Freight)</div> <div> July, 1929  <math>\frac{2408.3}{482} = 5.00 \text{ tons}</math> </div> <div> July, 1930  <math>\frac{2436.2}{482} = 5.05 \text{ tons}</math> </div> </div>					

The standardized averages are different from those calculated with the 1929 figures as standard, which were 5.05 tons, 5.11 tons; but the comparison between July, 1929, and July, 1930, still indicates a rise from 1929 to 1930, and that is the important point, because the crude average for all freight indicated a drop. Let us consider a further illustration of this method of obtaining standardized averages; this time taken from results of working in the Coal Mining Industry. The following figures refer to the six main coal producing districts in Great Britain, and the totals refer to operations in these six districts as a whole.

COAL MINING INDUSTRY: OUTPUT PER MAN-SHIFT (CWT)

	I Scotland	II Northumber- land	III Durham	IV South Wales	V Mid lands	VI Lanca- shire	Total
1924, 1st quarter	18.05	17.12	17.15	16.19	20.60	14.88	17.64
1925, 1st quarter	19.01	18.05	17.91	16.36	20.91	14.79	18.12

The output per man-shift varies from district to district. The average for all these districts in the first quarter of 1924 was 17.94 cwt., and in the first quarter of 1925 was 18.12 cwt. Increases in districts were recorded between these dates in four cases (I, II, III, IV) and decreases in the other two (V, VI). But between these two dates there was a change in the distribution of work done as the table below shows :—

MAN-SHIFTS WORKED (THOUSANDS)

		I	II	III	IV	V	VI	Total
1924, 1st quarter	Man-shifts	9 488	4 040	11 204	15,377	22,081	9 491	71 741
	Per cent	13.2	5.6	15.7	21.4	30.9	13.2	100
1925, 1st quarter	Man-shifts	8,495	3,366	9,416	14 025	22 213	8,632	66 147
	Per cent	12.8	5.1	14.2	21.2	33.6	13.1	100

In all these districts taken together there was a decline in the amount of work done, but in the case of the fifth district there was an increase. The percentage figures indicate the changes which have taken place district by district between the two periods. Now, since the output per man-shift is at different levels in different districts, the average for the whole is bound to be affected by the proportions of work done in these districts. Consequently any change in these proportions from one time to another will have an effect on the average figures, and therefore these average figures do not truly reflect changes due merely to organization of the industry. If we want to find out exactly a measure of such changes we certainly must eliminate the effect of changing distribution of work done and output between districts. This can be done by obtaining standardized averages, by referring the output per man-shift figures for each district to a standard distribution of labour, which will not be changed with

time. Let us take as standard the distribution of labour between districts which obtained during the whole year 1924 ; this is given by the table :—

YEAR 1924 MAN-SHIFTS WORKED (THOUSANDS)

I	II	III	IV	V	VI	Total
36,010	15,045	42,145	58,149	84,121	35,637	271,107

Make the calculations indicated in the table below :—

District	(1) Man-shifts in Standard period (thousands)	(2) Output per Man-shift 1924, 1st quarter (cwt )	(3) = (1) × (2)  Output mn cwt	(4) Output per Man-shift 1925, 1st quarter (cwt )	(5) = (1) × (4)  Output mn cwt.
I	36,010	18.95	682.5	19.01	684.6
II	15,045	17.12	257.6	18.05	271.5
III	42,145	17.15	722.3	17.91	754.8
IV	58,149	16.19	941.1	16.36	951.2
V	84,121	20.60	1732.8	20.30	1707.8
VI	35,637	14.88	530.1	14.79	527.0
	271,107		4866.4		4896.6

$$\begin{array}{l} \text{Standardized} \\ \text{Average Output} \\ \text{per man-shift} \end{array} = \frac{4866.4}{271,107} \text{ cwt} = 17.95 \text{ cwt}$$

$$\frac{4896.6}{271,107} \text{ cwt} = 18.06 \text{ cwt.}$$

Column (1) gives the standard distribution of work done, column (2) gives the district results in 1924, 1st quarter, column (3) gives the output which would be obtained if the amounts of work done were those in column (1), and the results of working were those in column (2). The total of column (3) would then be the total output produced by the total man-shifts given as

the sum of the figures in column (1), and the relationship of these is the output per man-shift for these districts as a whole, 17.95 cwt. A similar calculation on the working results in 1925, 1st quarter, shown in column (4), combined with the standard distribution of work done in column (1) gives the figures in column (5) showing the output which would be obtained under those circumstances. The average for the districts as a whole is 18.06 cwt. per man-shift. We may compare the crude averages with these standardized figures:—

	1924, 1st quarter	1925, 1st quarter
Crude	17.94 cwt	18.12 cwt
Standardized	17.95 "	18.06 "

These figures show that, on the whole, output per man-shift had increased between the two dates from 17.95 to 18.06 cwt. due to causes other than changes in the distribution of work done between the districts, an increase of 0.11 cwt., whereas the whole change in output from 17.94 to 18.12 cwt., i.e. an increase of 0.18 cwt., was partly due to the change in the work done in different districts. We could with reason say that 0.07 cwt. of this change is due to this last cause alone

This method of obtaining standardized ratios or averages which will serve for comparative purposes instead of the original or crude ratios or averages, enables us to separate the effects of different causes. This method is used when different areas are to be compared as to death-rates. As various areas have different age-distributions of the population, some places having perhaps

relatively more old persons or more infants, and as the death-rate changes with age, the differences in age-distribution from district to district themselves would cause differences between the death-rates, irrespective of whether there were differences between the death-rates in the same age-group from district to district. The effect of this cause on the death-rate is therefore removed by referring the death-rates of age-groups in each district to a standardized population age-distribution, and the standardized death-rates are worked out in the same way as has been indicated in the illustration above.

It is perhaps useful at this stage to interpolate some remarks on the place of ratios and averages, of the kind we have been discussing, in statistics. Statistics has been called the science of averages, and certainly these play a large part in any discussion of statistical results, because, as we have seen, they serve, in many cases, as the basis of the first simple kind of comparisons which can be made between different statistics. Especially is this the case when we are dealing with large numbers of units in our groups. But their importance must not be exaggerated. They may be considered as serving as rough guides to the information contained in the original figures from which they have been derived. Those interested in the efficient working of a huge organization like the railways look to such ratios and averages to serve as pointers of progress towards a greater efficient utilization of their resources, and if these figures, of which a great many are calculated, indicate changes which appear to be taking place and which are considered undesirable, the source of this change is sought in the original figures; a change in one of these ratios or

averages suggests a *prima facie* case for investigation, it does not necessarily mean that the change noticed in the ratio is connected with some loss of efficiency, there may be another and acceptable explanation. Similarly in the case of the Coal Mining Industry, the output per man-shift calculated at frequent intervals serves as an indication of progress in the industry. It is not supposed that every man-shift worked produces so much coal, there are many men employed in coal mines who are not working at the coal face, and if we wanted an indication of the work done by hewers we should work out another ratio altogether different, output per man-shift (hewers). But this average value will prove to be a guide to those interested in efficient working in this industry, and if any change occurs which suggests that coal is being produced at too expensive a rate of man-power, naturally investigation will be made to find out what changes in methods of production have caused this change in the ratio. In statistics relating to overcrowding in houses, an average of 2.33 persons per room in a household should not be considered in relation to any particular household: one should not quibble about its being impossible to consider 0.33 of a person,<sup>1</sup> but one should rather consider this average figure as giving a rough indication of circumstances relating to a whole group of households, and if such a figure is referred to a similar figure obtained from another group, say 1.24 persons per household, the difference between the sizes of these figures, which are representatives of each group, should call for comment and investigation into the different set of circumstances

<sup>1</sup> If we do not like to speak of 2.33 persons per household we can always think of 233 persons per hundred households, which will serve equally well

which give rise to these average results. These ratios or averages, which have been the subject of discussion in the previous pages, must be considered as broadly indicating the possession of a particular characteristic by a group of units, or the relation between two characteristics possessed by the group; they are not to be considered as having particular reference to any one individual of the group. Moreover, they are not to be considered as replacing the original information which is known relating to the group; it is possible to obtain more knowledge of a group than is indicated by such ratios or averages, by going back to the tables from which these are calculated.

We may consider that so far we have progressed by three stages. In the first we have the full detailed information which results when the inquiry is instituted; in the next stage, tabulation, a good deal of this detail is lost, but our information is available in a more concise form; in the next stage, when ratios and averages are obtained, still further detail is lost, and the information is available in a very simple form indeed. Or we may consider an analogy of this kind, the whole body of available information may be likened to a human body, the ratios and averages are the skeleton of this body, or they may be likened to the shadow of the body thrown by a light on a plane surface.

## CHAPTER 6

### THE CALCULATION OF AVERAGES

It sometimes happens that the figures in the original tables do not give us the information in such a form that ratios which are required are immediately calculable. Some intermediate arithmetical processes are necessary. For instance, consider the table below, which gives information relating to the number of births which have occurred in the first  $14\frac{1}{4}$  years of marriages, where husband and wife survived that length of time from the date of marriage, which were recorded in "Whitney, the descendents of John Whitney, who came from London, England, to Watertown, Massachusetts, in 1635". (F. C. Pierce, 1895) (This table is taken from *Biometrika*, XV, Parts 3 and 4. p. 415)

FREQUENCY OF OCCURRENCE OF BIRTHS IN  
(a) MARRIAGES BEFORE 1820, (b) MARRIAGES 1840-1859

Births	(a) Marriages before 1820		(b) Marriages 1840-1859	
	(1) Number of Marriages	(2) Number of Births	(3) Number of Marriages	(4) Number of Births
0	7	0	42	0
1	3	3	37	37
2	9	18	87	174
3	17	51	92	276
4	29	116	52	208
5	42	210	55	275
6	58	348	41	246
7	48	336	9	63
8	15	120	3	24
9	4	36	—	—
10	1	10	—	—
Total	233	1248	418	1303

In this table column (2) is obtained from column (1) by multiplying the figures in this column by the appropriate number of births. Similarly the figures in column (4) are obtained from those in column (3).

The appropriate ratios, the average number of births per marriage, are  $\frac{1248}{233} = 5.36$ ,  $\frac{1303}{418} = 3.12$ . In this table in its original form, where only columns (1) and (3) are shown, the total number of births, which is the numerator of the ratio, is not given, though the number of marriages is stated. The arithmetical calculations shown above have to be performed before we can get the ratio required.

Again consider such a table as this below, taken from 1921 Census Report.

DISTRIBUTION OF PRIVATE FAMILIES IN HUNSWORTH U D (YORKSHIRE)  
ACCORDING TO SIZE OF FAMILY AND NUMBER OF ROOMS

Number of Persons in Family	Number of Rooms												Number of Families
	1	2	3	4	5	6	7	8	9	10	11	12	
1	2	9	4	4	-	-	-	-	-	-	-	-	19
2	4	25	20	16	4	3	1	-	-	-	-	-	73
3	-	24	34	30	11	5	1	1	-	-	-	-	106
4	1	14	24	19	9	6	1	1	1	-	-	-	76
5	-	7	16	7	9	3	2	-	-	-	-	-	44
6	-	4	5	7	5	1	-	1	-	-	-	1	24
7	-	1	3	5	2	1	-	-	-	-	-	-	12
8	-	1	3	-	-	-	-	-	-	-	-	-	4
9	-	-	1	-	1	-	-	1	-	-	-	-	3
10	-	-	1	-	-	-	-	-	-	-	-	-	1
11	-	-	-	1	-	-	-	-	-	-	-	-	1
Number of Families	7	85	111	89	41	19	5	4	1	-	-	1	363

This table presents the information relating to the possession by the units of the group (private families)

Number of Persons in Family	Whole Group		Families with 2 Rooms		Families with 3 Rooms		Families with 4 Rooms		Families with 5 Rooms	
	Number of Families	Number of Persons	Number of Families	Number of Persons	Number of Families	Number of Persons	Number of Families	Number of Persons	Number of Families	Number of Persons
1	19	19	9	9	4	4	4	4	—	—
2	73	146	25	59	29	40	16	32	4	6
3	106	318	24	72	34	102	30	90	11	33
4	76	304	14	56	24	96	19	76	9	36
5	44	220	7	35	16	80	7	35	9	45
6	24	144	4	24	5	30	7	42	5	30
7	12	84	1	7	3	21	5	35	2	14
8	4	32	1	8	3	24	—	—	1	—
9	3	27	—	—	1	9	—	—	—	9
10	1	10	—	—	1	10	—	—	—	—
11	1	11	—	—	—	—	—	—	—	—
	363	1315	85	261	111	416	89	314	41	175
Average Persons per Family	1315 363	3.62	261 85	3.07	416 111	3.75	314 89	3.53	175 41	4.27

of two characteristics, the number of persons in the family, and the number of rooms available to the family, and at the same time it gives us the various ways in which one character is possessed by those units of the group which are alike in respect of their possession of the other character. Thus, we have the separate distributions of rooms in the cases of those families with 1, 2, 3, . . . persons in the family, and the separate distributions of persons in the family in the case of those families with 1, 2, 3, . . . rooms. There are therefore many interesting averages which may be calculated from the table on page 79.

We may, for instance, obtain, as shown in the above calculations, the average number of persons per family for the whole group, and for those families in the group with the same number of rooms. These results may be set out in this fashion —

	Whole Group	Families with			
		2 Rooms	3 Rooms	4 Rooms	5 Rooms
Average Persons per Family	3.62	3.07	3.75	3.53	4.27

From this we may see how the average number of persons per family (the average size of family) changes with the amount of accommodation (the number of rooms per family). Alternatively we may obtain the average number of rooms per family for the whole group, and for those families in the group containing the same number of persons. The calculations to this end are shown in the table on page 81.

Number of Rooms per Family	Whole Group		Families with 2 Persons		Families with 3 Persons		Families with 4 Persons		Families with 5 Persons	
	Number of Families	Number of Rooms	Number of Families	Number of Rooms	Number of Families	Number of Rooms	Number of Families	Number of Rooms	Number of Families	Number of Rooms
1	7	7	4	4	—	—	1	1	—	—
2	85	170	25	50	24	48	14	28	7	14
3	111	333	20	60	34	102	24	72	16	48
4	89	356	16	64	30	120	19	76	7	28
5	41	205	4	20	11	55	9	45	9	45
6	19	114	3	18	5	30	6	36	3	18
7	5	35	1	7	1	7	1	7	2	14
8	4	32	—	—	1	8	1	8	—	—
9	1	9	—	—	—	—	1	9	—	—
10	—	—	—	—	—	—	—	—	—	—
11	—	—	—	—	—	—	—	—	—	—
12	1	12	—	—	—	—	—	—	—	—
	363	1273	73	223	106	370	76	282	44	167
Average Rooms per Family	1273 363	3.51	223 73	3.05	370 106	3.49	282 76	3.71	167 44	3.80

These results may be shown in this way:—

	Whole Group	Families with			
		2 Persons	3 Persons	4 Persons	5 Persons
Average Rooms per Family	3.51	3.05	3.49	3.71	3.80

This summary of the results enables us to see how, on the average, the number of rooms per family increases with the size of family.

Moreover, it is possible also for us to construct from such a table as this, another table giving the distribution of these families according to the possession of another character "number of persons per room", obtained by dividing the number of persons by the number of rooms in the case of each family. If reference is made to the original table we can replace this by another (page 83), where the values of this third character are inserted.

We get the following distribution according to number of persons per room.

	0.25	0.33	0.45	1.15	1.45	1.75	2.05	2.35	2.65	2.95	3.25	3.55	3.85	Total
Number of Families	51	83	91	39	47	29	3	7	4	5	2	-	2	363

We may, alternatively, represent the distribution of families according to number of persons per room in a slightly different way, thus—

Number of Persons per Room	Less than 0.5	0.5-	1.0-	1.5-	2.0 and over	Total
Number of Families	19	115	130	53	46	363

DISTRIBUTION OF PRIVATE FAMILIES ACCORDING TO SIZE OF FAMILY, NUMBER OF ROOMS,  
AND NUMBER OF PERSONS PER ROOM

Number of Persons in Family	Number of Rooms											
	1	2	3	4	5	6	7	8	9	10	11	12
1	Persons per Room Number of Families	1 00 2	0 50 9	0 33 4	0 25 4	—	—	—	—	—	—	—
2	Persons per Room Number of Families	2 00 4	1 00 25	0 67 20	0 50 16	0 40 4	0 33 3	0 28 1	—	—	—	—
3	Persons per Room Number of Families	—	1 50 24	1 00 34	0 75 30	0 60 11	0 50 5	0 43 1	0 37 1	—	—	—
4	Persons per Room Number of Families	4 00 1	2 00 14	1 33 24	1 00 19	0 80 9	0 67 6	0 57 1	0 50 1	0 44 1	—	—
5	Persons per Room Number of Families	—	2 50 7	1 67 16	1 25 7	1 00 9	0 83 3	0 71 2	—	—	—	—
6	Persons per Room Number of Families	—	3 00 4	2 00 5	1 50 7	1 20 5	1 00 1	—	0 75 1	—	—	0 50 1
7	Persons per Room Number of Families	—	3 50 1	2 33 3	1 75 5	1 40 2	1 17 1	—	—	—	—	—
8	Persons per Room Number of Families	—	4 00 1	2 67 3	—	—	—	—	—	—	—	—
9	Persons per Room Number of Families	—	—	3 00 1	—	1 80 1	—	1 12 1	—	—	—	—
10	Persons per Room Number of Families	—	—	3 33 1	—	—	—	—	—	—	—	—
11	Persons per Room Number of Families	—	—	—	2 75 1	—	—	—	—	—	—	—

This table emphasizes the number of families where fairly large numbers of persons are living in houses with a comparatively small number of rooms. Thus there are 46 families out of 363 where the number of persons in the family related to the number of rooms is 2 or more.

Sometimes the information given in tabular form is such that it is impossible for us to determine the exact value of a ratio which we require, because we cannot get the numerator exactly, however many arithmetical processes we employ. Consider, for instance, such a table as this below, which gives the distribution of marks obtained by candidates in a certain examination, in a table such as is usually employed for this class of data.

Marks	35-	40-	45-	50-	55-	60-	65-	70-	75-	Total
Number of Candidates	1	5	12	34	23	22	23	12	1	133

In this kind of table, since the marks obtained by different candidates vary very widely, those are grouped together who get more or less the same marks, thus, there are 34 who get marks ranging from 50 to 54, there are 23 who get marks ranging from 55 to 59, and so on. With the loss of the original detail, which accompanies the condensation of the data into tabular form, we now have the difficulty that we cannot find the total marks obtained by all candidates, which should be the numerator of the ratio,  $\frac{\text{Total Marks obtained}}{\text{Total Candidates}} = \text{Average Mark}$ . But, for

many purposes, it is essential that we should be able to obtain such an average from a table of this kind, consequently we have to overcome this difficulty somehow,

and in practice we overcome it by means of a simple device which works well and which has a theoretical justification. We require the total number of marks of the whole 133 candidates; some of this total will be contributed by the 34 candidates, for instance, whose marks range from 50 to 54. Now we do not know exactly how much these contribute, but we know that this amount will be as much as, or more than  $34 \times 50 = 1700$ , and we know that it will be less than, or as much as  $34 \times 54 = 1836$ , because if the marks in this small group range from 50 to 54 they may be all at 50 or all at 54 or spread between 50 and 54. So that although we do not know the amount contributed exactly we know that it is between these limits 1700 and 1836. If we assume that the actual marks of those in this group are spread between 50 and 54 without being particularly concentrated at any particular number of marks in this range, then we shall not be far wrong if we assume that the average mark of those in this group is half-way between its limits 50 and 54, i.e. at 52 marks. In a table of this kind, this is a fair assumption which would only be unjustifiable if there was definite concentration of candidates at a special mark in the range, and if such concentration did exist then the original data would not be given to us in this particular tabled form. We must remember the origin of this kind of table. It is used in cases where so many and various are the values of the character possessed by the individuals in the group, that grouping together of those possessing the character in nearly the same manner is necessary. If the original data were such that the marks obtained by different candidates were the same in many cases, then this kind of table would not be necessary.

In such a case as that we would have a table of this kind :—

MARKS OBTAINED

	37	41	46	49	50	52	53	58	59	62	64	66	67	70	74	76	Total
Number of Candidates	1	5	5	7	10	9	15	11	12	8	14	13	10	7	5	1	133

We therefore assume that the average mark obtained by those candidates obtaining 50-54 marks is 52, and we get then an estimate of the contribution to the total marks of the whole group supplied by these 34 candidates, this is  $34 \times 52 = 1768$ . This assumption is used for each group, and we get, in this way, an estimate of the total marks obtained which we can use as the numerator of our ratio. As a matter of fact, this method of finding an estimate of this nature can be shown mathematically, on certain reasonable assumptions, to be justified, but the presentation of this argument is not suitable at this stage. We will therefore proceed to show the arithmetical work involved in the calculation of the estimated total number of marks obtained :—

Marks	Number of Candidates	Estimated Averages	Number of Marks
35-	1	37	37
40-	5	42	210
45-	12	47	564
50-	34	52	1768
55-	23	57	1311
60-	22	62	1364
65-	23	67	1541
70-	12	72	864
75-	1	77	77
	133		7736 Estimated Total Marks

$$\text{Average Mark} = \frac{7736}{133} = 58.2$$

# ORIGINAL MARKS

Columns A are marks Columns B are Number of Candidates

A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	Total
35		40		45	1	50	6	55	10	60	6	65	9	70	6	75				
38		41		46	1	51	4	56	3	61	6	66	6	71	3	76				
37		42	2	47	1	52	3	57	5	62	2	67	4	72	1	77	1			
38	1	43		48	4	53	7	58	4	63	7	68	4	73	2	78				
39		44	3	49	5	54	14	59	1	64	1	69		74		78				
35-	1	40-	5	45-	12	50-	34	55-	23	60-	22	65-	23	70-	12	75-	1			
Total Marks	38		218		575		1787		1294		1355		1521			851	77			7714
Estimates	37		210		584		1768		1311		1364		1541			864	77			7736
Error in Estimates	-1		-6		-11		-19		+17		+9		+20			+13	0			+22

Moreover, it may be pointed out that, unless there is definite bias in the distribution of the marks in the small groups, any error made in an estimate of the marks contributed to the total by one group will probably cancel with an error in the opposite sense made in respect of another group, so that the errors involved in these estimates are not cumulated when we get to the total estimate, which therefore should not involve a large error itself. Further, it must be remembered that when the average is calculated, this error is reduced considerably, seeing that it is divided by 133. The average then is likely to be very near the unknown correct result and can be trusted as a good approximation. It is interesting to test in this particular case the errors which are involved in these calculations. The original marks are easily referred to and may be shown in tabular form in the scheme on p. 87, where the contributions to the total marks are shown for each group and where the differences between the estimates and the correct figures are also shown, some of these are positive and some negative.

The correct average mark obtained from the original figures is  $\frac{7714}{133} \approx 58.0$ , the error in the estimate being 0.2 marks. This illustration may be considered as typical of the kind of error to which such an estimated average, calculated in this way, is liable. There is a further small point on method to which attention should be drawn. In the calculation of the estimated average some saving of arithmetical labour is possible by referring the marks to a new datum and by changing the mark-units. In the calculations worked in the table on p. 86,

if 37, 42, 47, etc., are rewritten in a different form, the arithmetic is simplified.

Marks	Number of Candidates	Marks	Estimated Number of Marks
37	1	$57 - 5 \times 4$	$1 \times (57 - 5 \times 4) = 1 \times 57 - 5 \times 4$
42	5	$57 - 5 \times 3$	$5 \times (57 - 5 \times 3) = 5 \times 57 - 5 \times 15$
47	12	$57 - 5 \times 2$	$12 \times (57 - 5 \times 2) = 12 \times 57 - 5 \times 24$
52	34	$57 - 5 \times 1$	$34 \times (57 - 5 \times 1) = 34 \times 57 - 5 \times 34$
57	23	57	$23 \times (57) = 23 \times 57$
62	22	$57 + 5 \times 1$	$22 \times (57 + 5 \times 1) = 22 \times 57 + 5 \times 22$
67	23	$57 + 5 \times 2$	$23 \times (57 + 5 \times 2) = 23 \times 57 + 5 \times 46$
72	12	$57 + 5 \times 3$	$12 \times (57 + 5 \times 3) = 12 \times 57 + 5 \times 36$
77	1	$57 + 5 \times 4$	$1 \times (57 + 5 \times 4) = 1 \times 57 + 5 \times 4$
	133		$133 \times 57 + 5 \times 108$
			$- 5 \times 77$
			Total Marks = $133 \times 57 + 5 \times 31$
Average = $\frac{133 \times 57 + 5 \times 31}{133} = 57 + \frac{5 \times 31}{133} = 57 + \frac{5 \times 0.233}{1} = 57 + 1.2 = 58.2$			

37 is written as  $57 - 20$  or  $57 - 5 \times 4$ , 42 as  $57 - 15$  or  $57 - 5 \times 3$ , and similarly with the others. In this way the arithmetic is considerably reduced, the result of course is the same. This process involves measuring the marks from a new datum 57 marks, instead of no marks, and also expressing the new marks in 5 marks units. In practice the scheme works as in table on p. 90.

By this simple device our multiplications are reduced to products by small numbers, 1, 2, 3, 4, 5, . . . The new datum is chosen at a convenient figure somewhere in the middle of the whole range of variations of the character, at a figure which is near those values of the character which are possessed by large numbers of the group. We try to replace the original measures of the variable

Marks	New Scheme	Number of Candidates	
35-	- 4	1	- 4
40-	- 3	5	- 15
45-	- 2	12	- 24
50-	- 1	34	- 34
55-	0	23	- 77
60-	1	22	22
65-	2	23	46
70-	3	12	36
75-	4	1	4
		133	108

Total + 31

Average + .233.

Average in original units and  
scales =  $57 + 5 (.233) = 58.2$

by  $\pm 1$ ,  $\pm 2$ ,  $\pm 3$ , . . . and at the same time to associate the smallest of these with the largest of the numbers which have to be multiplied. This insistence on reducing the arithmetical labour involved in the calculations is not altogether imposed because we do not want to do too many computations, but also because the simpler the calculations the less the liability is there to error in our results, and above all, we want accuracy. A further illustration (p. 91) should make the above points clear.

In this table, the large numbers in the distribution are in the range 20 cm. to 30 cm., and we take the new datum at the middle of the range 24-26 cm.

A further difficulty sometimes arises in the calculation of the average in those cases where the information in the tables lacks exactness. In the table below, for instance,

DISTRIBUTION OF LENGTH OF PLAICE MEASURED IN 1907  
*From Journal of Royal Statistical Society, 1925, p 245*

Length (cm.)	New Units and Scale	Number of Plaice	
18-	- 3	12	- 36
20-	- 2	101	- 202
22-	- 1	258	- 258
24-	0	297	- 496
26-	1	152	152
28-	2	80	160
30-	3	47	141
32-	4	22	88
34-	5	10	50
36-	6	6	36
38-	7	5	35
40-	8	4	32
42-	9	2	18
44-	10	2	20
		998	732

Total + 236      Average +  $\frac{236}{998} = + 23647$  (new scale and units).

Average length (original scale and units)  
 $= 25 + 2 \cdot (23647) \text{ cm} = 25.47 \text{ cm}$

taken from the Report of the Royal Commission in the Coal Industry 1925, Annex, p. 263, which gives information on the profitability of undertakings, this difficulty arises.

UNDERTAKINGS IN THE COAL MINING INDUSTRY, GREAT BRITAIN, 1923,  
 SHOWING PROFIT OR LOSS PER TON

Profit or Loss per ton	Loss				Under 1s loss or profit	Profit				Total
	7s and over	5s -	3s -	1s -		1s -	3s -	5s -	7s and over	
Number of Undertakings	10	15	24	65	166	203	120	32	12	653

If we wish to calculate the average profit or loss per ton for the whole group, we can make the same kind of assumptions as before as to the estimated average profit or loss in the case of each small group of undertakings in the table, except in respect of the first and last. We were guided before by the limits of the ranges and assumed the average to be in the middle, but we obviously cannot do this when the limits of a group are not both defined. What is the estimated average loss per ton of the 16 undertakings in a group described as 7s. and over, what is the estimated average profit per ton of the 12 undertakings in the group with a profit of 7s. and over? We do not know whether the greatest loss incurred in a particular case is 9s., 10s., or 15s.; similarly we do not know the upper limit to which profits extended. Consequently, any assumption we make

Profit or Loss		New Scale and Datum	Numbers of Undertakings	
Loss	7s. and over	- 4	16	- 64
	5s. -	- 3	15	- 45
	3s. -	- 2	24	- 48
	1s. -	- 1	65	- 65
UNDER	1s. loss			
	1s. profit	0	166	- 222
PROFIT	1s. -	1	203	203
	3s. -	2	120	240
	5s. -	3	32	96
	7s. and over	4	12	48
			653	+ 587
				Total + 365

$$\text{Average} + \frac{365}{653} = 559$$

$$\text{Average profit, original scale: } 0s. + 2(-56)s. = 1.12s$$

as to the contribution from these two classes to the total profit or loss for all undertakings will have a further disturbing influence on this total. In practice, if we cannot have access to the original data so that more precise knowledge may be gained, we would base these estimates on the assumption that there was the same interval between the limits of these extreme groups as in the case of the other groups where the limits are known. Thus we should make our calculations as shown in table on p. 92.

It is interesting to find out the difference between this average and that calculated on some other assumption. Let us suppose, for instance, that the losses and profits in the extreme cases ranged up to 11s in each, then our calculations would be shown in this way—

Profit or Loss		New Scale and Datum	Number of Undertakings	
Loss	7s and over	- 4½	16	- 72
	5s -	- 3	15	- 45
	3s -	- 2	24	- 48
	1s -	- 1	65	- 65
UNDER	1s loss	0	166	- 230
	1s. profit			
PROFIT	1s.-	1	203	203
	3s -	2	120	240
	5s.-	3	32	96
	7s and over	4½	12	54
			653	+ 593
			Total + 363	

$$\text{Average} + \frac{363}{653} = .555$$

$$\text{Average profit, original scale: } 0s. + 2(.555)s = 1.11s.$$

The change in average profit per ton is one hundredth part of a shilling, about one eighth of a penny, a negligible amount in this connection. There are two factors which enter into the calculations which reduce this difficulty in cases of this sort to negligible proportions. The first is that errors involved in assumptions of the averages of the extreme groups will tend to balance one another in the final total,<sup>1</sup> and the other is that the numbers involved in this kind of table are small compared with the total numbers, so that any error introduced is reduced considerably when the numerator is divided by the denominator. Of course, if the number of cases at the extremes of a table like this were large compared with the other numbers in the table, this method of procedure could not be adopted because the errors introduced then might be considerable, but the fact is, that if these numbers were large, more precise information would be given in the original table about them. It is only because the numbers in this kind of table at the extremes are small that they are considered as rather of less importance than the others, and the compilers of the table dismiss them with, "there is a small group of 12 out of 653 with a profit of 7s. and over, and it is not worth while to specify exactly what these individual profits are."

Finally, let us consider the table on p. 95 relating to numbers of incomes assessed for super-tax, and the amount of these incomes.

The original table from which these figures were extracted gave the columns (1), (2), and (3). Column (4) is obtained by dividing the figures of column (3) by the

<sup>1</sup> We do not mean by this that there is in all such cases complete cancellation

U.K. DISTRIBUTION OF INCOMES LIABLE TO SUPER-TAX  
1924-5

(1) Income (£000)	(2) Number of Persons	(3) Total of Incomes assessed (£000)	(4) Average Income (£000)	(5) Middle Income of range (£000)
2-	23,225	51,964	2.14	2.25
2.5-	15,493	42,358	2.73	2.75
3-	18,385	63,242	3.44	3.5
4-	10,294	45,909	4.45	4.5
5-	6,382	34,755	5.45	5.5
6-	4,303	27,818	6.47	6.5
7-	3,051	22,817	7.47	7.5
8-	3,957	35,414	8.95	9.0
10-	4,606	55,687	12.2	12.5
15-	1,970	33,929	17.4	17.5
20-	1,025	22,728	22.1	22.5
25-	554	15,087	27.2	27.5
30-	589	20,231	34.4	35.0
40-	304	13,532	44.5	45.0
50-	327	19,475	59.5	62.5
75-	128	10,977	85.6	87.5
100-	143	28,809	201.5	—
Total	94,736	544,732	5.75	—

corresponding figures of column (2). Column (5) is obtained from column (1). It is necessary to note that the figures in column (4) are all less than the corresponding figures of column (5). Now if the whole of the information available to us were contained in columns (1) and (2), and if we proceeded to the estimate of the total income assessed for this purpose by using column (5), and multiplying these figures by those in column (2), and then adding the results, the estimated average would definitely be in excess of its true value, because all the figures in column (5) are greater than the true averages for each income grade. The method described above for obtaining an estimate of the average of a group which shall be close

to the real figure definitely breaks down in this case. This method can only be used with safety when the distribution of the numbers in the group, according to their possession of a given characteristic, is such that only small numbers of the group have the largest and smallest values of the characteristic, while the main body of the group possess the character to a moderate extent. Thus, in the table showing the distribution of marks of 133 candidates, only 6 had less than 45 marks, only 13 had more than 70 marks, the remainder had marks between 45 and 70 marks. In such a table the numbers in the group, when distributed according to the values of a particular characteristic, increase from zero to a maximum and then decrease again to zero. Now in the table of those liable to super-tax, the largest group of persons is that with incomes above £2,000 but less than £2,500, the first group in the table, thereafter the numbers gradually diminish. This is an entirely different kind of distribution from the others already considered, and with this type the assumptions we made are not justified, and the method described of calculating the average, cannot be used. The difference between these types of distribution is very evident when they are presented in graphical form.

## CHAPTER 7

### GRAPHICAL METHODS

Statistical data which are available in the form of tables are represented graphically with the object of making possible the appropriate comparisons more easily. Most people can normally appreciate the relative sizes of a number of figures more readily when these are pictorially presented, than they can by looking at a table. As an aid therefore to a proper realization of the relative sizes of different numbers, graphical methods are extremely useful. In some cases, of course, a graphical representation may merely give a rough approximation to the actual figures, this depends on the scale which is used. For instance, it is difficult to distinguish between 313 and 314 when represented on a scale where 100 is equal to 1 in., the difference might exist and be important, but it would not be noticeable in such a diagram.

Diagrams which are used to show the numbers of units in a group possessing different values of a character are of two kinds. In the first kind, which is illustrated in *Diagram 1*, showing the number of families with different numbers of persons in the family (from the table p. 78), the numbers of persons per family are represented on a convenient horizontal scale, and the number of families corresponding are represented vertically by means of separate rectangular blocks of the same thickness. The number of families is indicated in each case by the height

of the appropriate rectangular block, which is given thickness merely in order to improve the appearance of the diagram. The scale on which the number of families is represented is a linear scale and, of course, a line must be given thickness, if it is to be visible, and we go further in practice and instead of a line, use a rectangle. In some cases this rectangle is nothing more than a line of

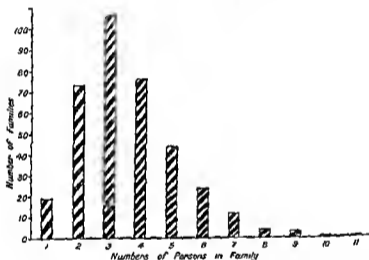


DIAGRAM 1 1921 Hansworth UD Distribution of 363 Private Families, according to number of persons in Family

some considerable thickness. This practice is general in all cases of this kind where the variable character can be counted. We differentiate between the different numbers of the character by separating them on the horizontal scale by sufficient space, so that neighbouring rectangular blocks or thick lines are absolutely distinct from each other.

In the second kind, where the variable character is measurable, it is represented again on a horizontal scale, but the number of units in the group possessing the character between certain limits is represented graphically by means of areas. The scale used for the numbers is an area scale. Since the number which we wish to represent graphically is a number of units possessing the character

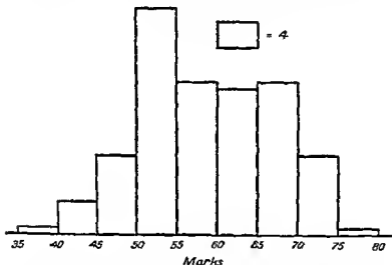


DIAGRAM 2 Distribution of Marks of 133 candidates

between certain limits, it is reasonable to do this by means of a figure standing on a base corresponding to this interval. The simplest figure of this kind is a rectangle, and the area of the rectangle corresponds to the number of cases having values of the character coming within the range of that interval. Thus Diagram 2 gives a graphical representation of the table on p. 84. The areas of the rectangles correspond to the numbers of

candidates with marks between the various limits shown in the table. The rectangles in this kind of diagram are contiguous, and this is reasonable, since the numbers in the original table merge into one another in the sense that there will be certain individuals with 49 marks (say) who should be represented in a diagram in close proximity

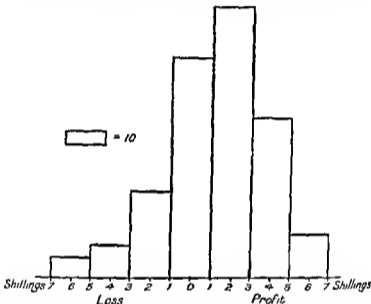


DIAGRAM 3 Coal Mining Industry G B 1923 Distribution of Undertakings by Profit or Loss

to individuals with 50 marks. Diagram 3 illustrates a difficulty which is encountered in this kind of graph. In the original table the two extreme groups are merely indicated by the vague description "loss of 7s and over", "profit of 7s and over". The vagueness of description gives rise to a perplexity when we wish to make a graphical

representation. We do not know what are the bases of the rectangles which should be constructed to correspond to the numbers in these two groups; all we know is the area. Consequently, we cannot show in the diagram that part of the table containing these two end groups; we therefore exclude them from the diagram. We might, of course, assume that "and over" in each case means 9s., and erect our rectangles on the base corresponding to

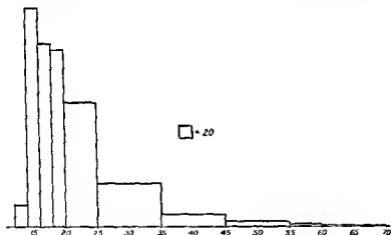


DIAGRAM 4 Warrington Ages of Textile Workers (Female)

7s. to 9s. But equally we might assume that "and over" means up to 10s. or 11s. and get a slightly different appearance in the resulting diagrams. Not much is lost by leaving these out.

Diagram 4 illustrates a different point. In some cases, as in the table below, the limits of the class-intervals of the variable character are not at regular intervals in the scale.

## CENSUS 1921. WARRINGTON

*Occupied Females 12 years old and over**Age-distribution of those engaged in Textile Occupations*

Age	12-	14-	16-	18-	20-	25-	35-	45-	55-	60-	65-69	Total
Number	34	346	290	280	402	344	98	43	10	6	4	1947

The graphical representation of such a table consists of a number of rectangles of different widths corresponding to the changing age-limits of the grades.

Great care must be taken over the construction of these

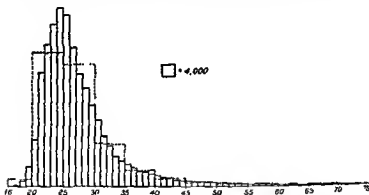


DIAGRAM 5 England and Wales Age-distribution of Bachelors who married in 1925

diagrams. When the grade intervals are the same, the rectangles stand on equal bases, and the areas are proportional to the heights of these rectangles, thus for practical purposes a linear scale is used for plotting. But when the grade intervals are not the same, the persons constructing the diagram must think in terms of areas and not lengths, when the appropriate rectangles are being drawn.

Diagram 5 again brings out the use of areas to represent

numbers of units. This diagram shows graphically the table below giving the age-distribution of bachelors in *England and Wales* who married in 1925. This table gives the numbers at each year of age, and the numbers in 5-year groups. In the diagram the area of each rectangle on the wider bases is equal to the total of the areas of the corresponding rectangles on the narrower bases (1 year).

ENGLAND AND WALES: AGES OF BACHELORS WHO MARRIED IN 1925

Age	Number	Age	Number	Age	Number
16	14	39	1,963	62	65
17	106	40	1,710	63	60
18	971	41	1,311	64	47
19	3,320	42	1,297	65	64
20	7,596	43	968	66	31
21	18,698	44	928	67	27
22	23,233	45	867	68	30
23	26,538	46	659	69	35
24	29,129	47	594	70	24
25	28,024	48	497	71	16
26	22,605	49	483	72	9
27	19,243	50	436	73	10
28	16,022	51	331	74	13
29	13,206	52	311	75	2
30	10,828	53	233	76	2
31	8,209	54	215	77	2
32	6,629	55	196	78	—
33	4,987	56	167	79	1
34	4,245	57	140	80	1
35	3,627	58	134	81	—
36	3,096	59	105	82	1
37	2,501	60	122	83	1
38	2,380	61	89		

DISTRIBUTION IN AGE GROUPS

Age	Under 20	20-	25-	30-	35-	40-	45-	50-	55-	60-	65-	70-	75-
Number	4,413	105,194	99,300	94,898	13,567	6,214	3,100	1,526	742	383	187	72	1

In the age-group 30-34 there are altogether 34,898 bachelors, of whom 10,828 are 30 years old, 8,209 are 31, 6,629 are 32, 4,987 are 33, 4,245 are 34; and since the large difference between the numbers in successive 5-year groups, e.g. 34,898 in the group 30-34 and 13,567 in the group 35-39, is due to the grouping (we note that there are 3,627 at 35, 3,096 at 36, and so on, numbers which are of the same order as 4,987 at 33 and 4,245 at 34), it would perhaps be preferable if, instead of making the graph with discontinuous

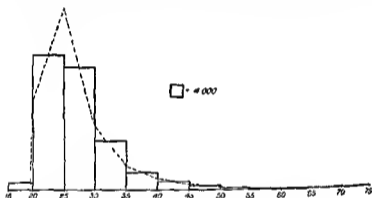


DIAGRAM 6 England and Wales Age-distribution of Bachelors who married in 1925

"steps" which are seen when rectangular blocks are used, we employed trapezia which would give a continuous outline to the graph. This is done in the hope that the graphical representation thus obtained will be a better approximation to the original distribution than that when rectangles are used. This suggestion is followed up in the construction of Diagram 6, where the same data are again shown, both by rectangles and trapezia. If this diagram is compared with 5 it will be seen that the outline made

by the sloping sides of the trapezia conforms to that made by the rectangles when the numbers at each age are plotted. This polygon diagram is, in fact, preferable to the rectangular block diagram (sometimes called a *histogram*), but it is normally more troublesome to draw, and the histogram type of graph is generally used, always understanding that this is really a crude approximation to a shape, to which the polygon is a better approximation.

The polygon graph is constructed in the following manner. Let us suppose that we are dealing with the graphical representation of a table such as this, where the histogram would show rectangles increasing in height and then decreasing again as  $x$  increases.

Variable character	$x-$	$(x + h)-$	$(x + 2h)-$	. . . . .	$(x + (m - 1)h)-$
Number	$n_0$	$n_1$	$n_2$		$n_{m-1}$

The widths of the trapezia are  $h$  units, suppose this corresponds to  $h$  inches on the horizontal scale. We start the graph with a triangle (instead of a trapezium) at one end of the scale. If one unit is to be represented by  $kl$  square inches, then if  $n_{m-1}$  (say) is represented by a triangle of  $n_{m-1}kl$  sq. in. on a base of  $h$  in., the vertical side of the triangle is  $2n_{m-1}l$  in. Then  $n_{m-2}$  is represented by a trapezium of  $n_{m-2}kl$  sq. in. in area, on a base  $h$  in., one vertical side of which has just been determined as  $2n_{m-1}l$  in., and the other will necessarily be  $(2n_{m-2} - 2n_{m-1})l$  in. The next number  $n_{m-3}$  is represented by a trapezium of area  $n_{m-3}kl$  sq. in., on a base of  $h$  in., one vertical side is already determined as  $(2n_{m-2} - 2n_{m-1})l$  in., the other is therefore  $(2n_{m-3} - 2n_{m-2} + 2n_{m-1})l$  in.

Subsequent trapezia are constructed in like manner. The trapezium corresponding to  $n_s$  on a base from  $x + sh$  to  $x + (s + 1)h$  would have the vertical side at  $x + sh$  of length  $2(n_s - n_{s+1} + n_{s+2} \dots + (-1)^{m-1-s} n_{m-1})l$  in, and the other vertical side at  $x + (s + 1)h$  of length  $2(n_{s+1} - n_{s+2} + n_{s+3} \dots + (-1)^{m-2-s} n_{m-1})l$  in. The last trapezium (on a base  $x$  to  $x + h$ ) would be a triangle if  $2(n_0 - n_1 + n_2 \dots + (-1)^{m-1} n_{m-1})l$  in. is equal to zero. This is not usually the case, for instance, in the illustration used, where  $n_0 = 4413$ ,  $n_1 = 109,607$ ,  $n_2 = 99,300$ , etc.  $n_0 - n_1 + n_2 \dots = -31,387$ , and in such cases we must either have a trapezium instead of a triangle, or else have a triangle on a smaller base. This latter alternative is necessary in our illustration. Here the vertical side of the trapezium at 20 years is  $2(35,800)l$  in. The area of the triangle of which this is to be the vertical side is  $4,413kl$  sq. in., the length of the base is therefore  $\frac{4413}{35800} k$  in., and since 5 years is represented by  $k$  in., this corresponds to about one eighth of 5 years, i.e. just over one half of a year.

When the length of the class interval changes the procedure is not so easily described in general terms, but an illustration will suffice to show what is done. Suppose we wished to represent in this way the following table.

DISTRICTS IN ENGLAND AND WALES HAVING LESS THAN 5,000 INSURED WORKERS, DISTRIBUTED ACCORDING TO UNEMPLOYMENT PERCENTAGE, AT 17TH SEPTEMBER, 1928

Percentage Unemployment	Under 2	2	4-	6-	8-	10-	15-	20-	30-	40-	Total
Number of districts	17	40	37	26	20	38	21	15	13	8	236

Take as scale of unemployment percentage, 10 per cent equal to 1 in., and 50 districts equal to 1 sq. in. Start with the 17 districts with percentage greater than 0, but less than 2, the area to correspond will be 0.34 sq. in., this will be represented by a triangle on a base of 0.2 in. and height 3.4 in. The next number 40 will be represented by the area 0.80 of a trapezium on base 0.2 in., one vertical side being 3.4 in., the other 4.6 in., and so on. We can show the results of these calculations in the table below:—

Percentage Unemployment	Number	Area Sq in	Base In	Height of sides
Under 2	17	.34	.2	0
2-	40	.80	.2	3.4
4-	37	.74	.2	4.6
6-	26	.52	.2	2.8
8-	20	.40	.2	2.4
10-	38	.76	.5	1.8
15-	21	.42	.5	1.44
20-	16	.32	1.0	.24
30-	13	.26	1.0	.40
40-	8	.16	2.67	.12
				0

The last calculation concerning the base of a triangle of area 0.16 sq. in. and height 0.12 in. gives 2.67 in. Thus the polygon ends at 66.7 on the scale of percentages, whereas actually if we consult the original figures we find that the highest percentage unemployment of a

district was 69.5. The graph drawn from these calculations is shown in Diagram 7.

The distribution of incomes liable to super-tax shown in the table on p. 95 is represented graphically in Diagram 8.

In all these diagrams it is to be observed that when there are great contrasts between the sizes of some of the

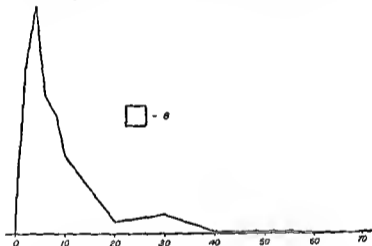


DIAGRAM 7 England and Wales 1928 (17th September) Distribution of Districts according to Unemployment Percentage

figures to be plotted, the smallest are so small that they cannot be shown on the graph on the scale chosen, which is determined by the range of figures in the table and the limited size of the paper at our disposal

### *Cumulative Diagrams*

There is another method of graphing such tables as these which we have been considering. This new method

gives, as a matter of fact, a *graphical* representation of another table obtained from the original one. This table is called a cumulative table and it shows numbers of units

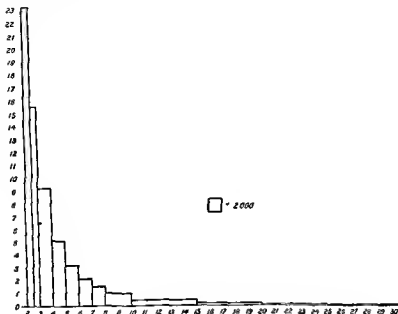


DIAGRAM 8 United Kingdom, 1924-5 Distribution of Incomes liable to Super-tax

in the group having less or more than successive values of the measurable character. Thus the table relating to

Percentage Unemployment	Under 2	2-	4-	6-	8-	10-	15-	20-	30-	40-	Total
Number of districts	17	40	37	26	20	38	21	16	13	8	230

Unemployment Percentages in different districts, might be replaced by a table of this kind.—

A CUMULATIVE TABLE SHOWING THE NUMBER OF DISTRICTS WITH UNEMPLOYMENT PERCENTAGE BELOW CERTAIN LIMITS

Percentage Unemployment below	2	4	6	8	10	15	20	30	40	70
Number of districts	17	57	94	120	140	178	199	215	228	236

Alternatively, the table might be replaced by a table of this kind :—

A CUMULATIVE TABLE SHOWING THE NUMBER OF DISTRICTS WITH UNEMPLOYMENT PERCENTAGE ABOVE CERTAIN LIMITS

Percentage Unemployment above	0	2	4	6	8	10	15	20	30	40
Number of districts	236	219	179	142	116	96	58	37	21	8

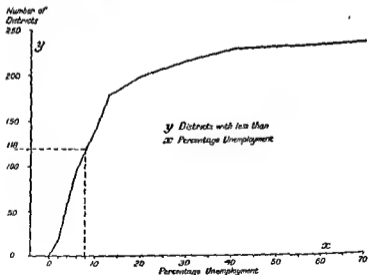


DIAGRAM 9 Cumulative Diagram.

The figures in such cumulative tables are plotted to make a cumulative diagram, by representing the numbers on a linear scale. The variable quantity of which we have the measurements (in this case Unemployment Percentage) is scaled horizontally, the scale of numbers is shown vertically. Points are plotted corresponding

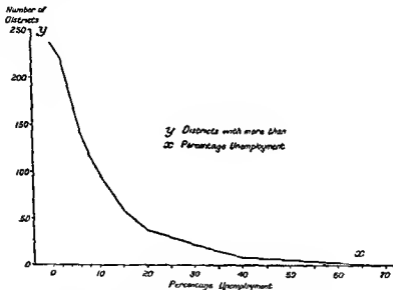


DIAGRAM 10 Cumulative Diagram.

to these pairs of values 2, 17; 4, 57; 6, 94; and so on, in the one case, and to 0, 236; 2, 219; 4, 179; and so on, in the other case. These points are joined together to make a polygon outline as in Diagrams 9 and 10. It must be emphasized that in this kind of diagram the numbers are now plotted on a linear scale, and not on an area scale as heretofore. The cumulative diagrams can

(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
0.3	2	2	6.9	4	107	16.1	1	181
0.7	1	3	7.0	2	109	16.5	2	183
1.1	1	4	7.1	1	110	16.6	2	185
1.2	2	6	7.2	4	114	16.7	1	186
1.4	1	7	7.4	1	115	16.9	1	187
1.5	1	8	7.5	1	116	17.1	2	189
1.6	3	11	7.6	1	117	17.4	1	190
1.7	2	13	7.8	1	118	18.4	2	192
1.8	2	15	7.9	2	120	18.5	1	193
1.9	2	17	8.0	2	122	18.9	1	194
2.0	1	18	8.1	1	123	19.0	2	196
2.1	1	19	8.2	1	124	19.2	2	198
2.2	1	20	8.4	2	126	19.9	1	199
2.3	6	26	8.5	2	128	20.2	1	200
2.4	6	32	8.7	2	130	20.5	1	201
2.5	3	35	8.9	1	131	21.5	1	202
2.7	1	36	9.0	1	132	21.7	1	203
2.9	1	37	9.2	1	133	22.0	1	204
3.0	1	38	9.3	1	134	22.5	1	205
3.1	3	41	9.7	4	138	24.3	1	206
3.2	2	43	9.8	1	139	24.5	1	207
3.3	4	47	9.9	1	140	25.0	1	208
3.4	1	48	10.0	1	141	25.2	1	209
3.5	2	50	10.1	2	143	25.6	2	211
3.6	2	52	10.2	1	144	26.0	1	212
3.7	3	55	10.4	1	145	27.0	1	213
3.8	1	56	10.5	2	147	28.6	1	214
3.9	1	57	10.6	2	149	29.0	1	215
4.0	2	59	10.7	1	150	30.1	1	216
4.3	2	61	11.0	1	151	30.5	1	217
4.4	6	67	11.3	1	152	31.2	1	218
4.5	6	73	11.4	1	153	31.4	1	219
4.6	1	74	11.6	3	156	31.8	1	220
4.7	3	77	11.9	1	157	33.1	1	221
4.8	1	78	12.0	2	159	33.5	1	222
4.9	1	79	12.3	1	160	34.7	1	223
5.0	2	81	12.5	2	162	37.1	1	224
5.1	4	85	12.7	1	163	37.8	1	225
5.2	1	86	12.8	1	164	38.6	1	226
5.4	1	87	13.4	1	165	38.8	1	227
5.5	3	90	13.6	1	166	39.0	1	228
5.7	1	91	13.8	2	168	41.1	1	229
5.8	2	93	13.9	2	170	43.3	1	230
5.9	1	94	14.1	3	173	48.7	1	231
6.0	1	95	14.2	2	175	53.5	1	232
6.1	2	97	14.5	1	176	55.7	1	233
6.2	1	98	14.6	1	177	58.6	1	234
6.4	2	100	14.7	1	178	62.2	1	235
6.5	1	101	15.4	1	179	69.5	1	236
6.7	2	103	16.0	1	180			

(1) = Unemployment percentage (2) = Number of districts

(3) = Summation

be used to determine the number of districts with less or more than a certain percentage of unemployment not specified in the table. This number would correspond in the diagram to the vertical ordinate ( $y$ ) erected at the appropriate point ( $x$ ) on the horizontal scale. For instance, in Diagram 9 we can read off when  $x = 12$ ,  $y = 155$ , and deduce that roughly 155 districts had an unemployment percentage less than 12. Naturally, any result of this kind will be only approximate. Alternatively, for a given  $y$  we can find the corresponding  $x$ . Thus when  $y = 118$ ,  $x = 7.9$ , and we can say that half the districts (118 out of 236) had less than 7.9 per cent unemployed, and the other half had more than this percentage of unemployment.

A cumulative diagram is really an approximate graphical representation of the group of values of the character when these are arranged in order of size and the number of cases summed successively. The table on p 112, giving the individual district unemployment percentages, and the results of successive summation will make this clear.

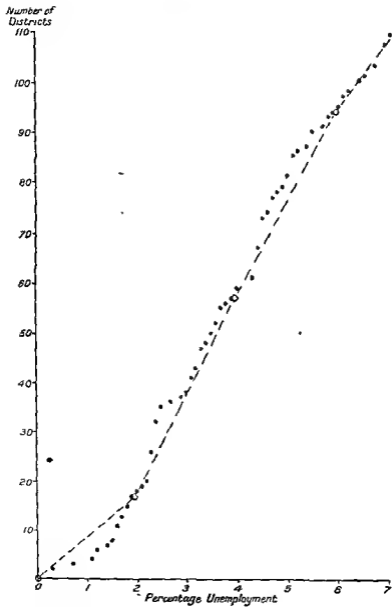
In this table it should be observed that the results of the summation corresponding to the different unemployment percentages do not tally exactly with the corresponding figures of the table from which Diagram 9 is constructed. Thus in the complete summation table above there are 17 districts with a percentage 1.9 or less, 57 districts with 3.9 or less, and so on, whereas in the cumulative table on p. 110 there are shown 17 districts with less than 2 per cent unemployment, 57 districts with less than 4 per cent unemployment, and so on. This apparent incompatibility can be shown to be illusory if we realize that each percentage has been worked out correct to the first

decimal figure only, and that a percentage of 3.9, for instance, may represent any number from 3.85 to 3.95, and a percentage of 4.0 may stand for any figure between 3.95 and 4.05. Consequently, we ought really, in our cumulative table on p. 110, to change the unemployment percentages shown there, and present the table in the form :—

Percentage under	1 95	3 95	5 95	7 95	9 95	14 95	19 95	29 95	39 95	69 95
Number of districts	17	57	94	120	140	178	199	215	226	236

The corresponding information in the complete cumulative table (p. 112) would then read, for instance, instead of 17 districts with a percentage of 1.9 or less, 17 districts with less than 1.95 percentage, and so on. From the point of view of the graphical representation, this change in the  $x$  unit has such a slight effect that it is hardly noticeable in a diagram, and in practice the table on p. 110 is used as the basis of the graph instead of the table above.

Diagram 11 shows the beginning of the cumulative graph when the individual percentages are known. It is constructed from the data of the table on p. 112 by plotting the results of successive summation against the corresponding unemployment percentage. In reality, Diagram 9 is an approximation to Diagram 11. This last gives the true facts relating to the percentages in different districts, but the table on p. 110, which is the kind of table which as a rule is available to us, gives only a summary of the total information about the districts. This table, then, supplies us with only a few of the points of our real cumulative diagram; these are indicated



by small circles in Diagram 11. Straight lines joining them will give us approximately the position of the unknown part of the cumulative graph. In this way, we get Diagram 9 as an approximation to Diagram 11.

A cumulative diagram shows us then, by means of vertical ordinates, the approximate number of the whole group possessing less than (or more than) any measure of the character represented on the horizontal scale. In effect this kind of diagram puts in order of magnitude the individuals of the group, and is very helpful in assisting us to pick out measurements possessed by certain individuals.

One of these of most value is the *Median* measurement, which is used as a representative of the group.

## CHAPTER 8

### THE MEDIAN AND MEASURES OF DISPERSION

We have described how the average is used for purposes of comparison ; the average wage of a group of persons is perhaps used to represent the group in a comparison with another group. Similarly the median is used. The median is an obvious measurement to take ; it is that of the middle individual in the group, when those in the group are disposed in order according to the size of their measurements of the particular character we are considering. Moreover, if all the detailed measurements of the group are known, the median is easily identified without any calculation. Thus, if there are seven persons of heights 5 ft. 7 in., 5 ft. 9 in., 5 ft. 10 in., 5 ft. 11 in., 6 ft., 6 ft., 6 ft. 1 in., the median is 5 ft. 11 in., the height of the middle person, when all are arranged in order of height. In order to get the average we should have to calculate the total height 41 ft. 2 in. and divide by 7, giving 5 ft. 10½ in. It is seen that the median and average are not necessarily the same, though in certain cases they may be identical or approximately identical. If there is no middle individual, as happens when there is an even number in the whole group, the median is taken as the average of the two middle measurements. Thus, the median of the following group of wages 45s., 46s. 6d., 47s., 48s., 48s. 6d., 49s., 49s., 50s., is 48s. 3d., half-way between the two middle wages, 48s. and 48s. 6d. Generally,

then, practically half of a total group of individuals possess the character to an extent less than the median measurement, and the other half possess the character to an extent greater than the median. Now we see how the cumulative diagram can be used to identify the value of the median, when there are large numbers in the group. The ordinates of such a diagram, corresponding to values of the character on the horizontal scale, indicate the numbers in the group with less (or more) than those values of the characteristic. Consequently, the vertical ordinate corresponding to the median value in the horizontal scale will be exactly half the total number in the group. Thus, in Diagram 9, the vertical ordinate indicating half of the total 236 corresponds to 7.9 in the horizontal scale. The median percentage is 7.9. The complete details on p 112 show that the 118th district has a percentage of 7.8, the 119th and 120th districts have percentages 7.9, thus more accurately the median is 7.85. Diagram 9 is, as we pointed out, only approximately correct, consequently any estimate of the median, obtained from it, is itself necessarily only approximate.

The median can also be obtained by calculation, without drawing the cumulative graph, by reference to the cumulative table itself. Thus, in the case of our illustration, the table on p 114 shows 94 districts with percentages less than 5.95 and 120 with less than 7.95; the middle 118th and 119th will be somewhere in this group of 26 with percentages between 5.95 and 7.95. Assuming that the percentages of these 26 districts are evenly distributed between these limits, by a simple proportion sum we identify the median as 5.95

$+\frac{118\frac{1}{2}-94}{26} \times 2$ , which is  $5.95 + \frac{24\frac{1}{2}}{26} \times 2$ , giving 7.8 to the nearest decimal place.

Let us consider another example. On p. 91, there were shown figures relating to profit and loss sustained by undertakings in the Coal Mining Industry. Without constructing the whole cumulative table we can write down the information we need in this form :—

<i>Number of Undertakings.</i>	<i>Profit less than</i>
286	1/-
489	3/-
<hr/> Total 653	

The middle undertaking is the 327th, which will be in the group of 203 with more than 1s. but less than 3s. profit. Assuming that the profits of these 203 are evenly distributed in the range from 1s. to 3s., we identify the median as 1s.  $+\frac{327-286}{203} \times 2s.$ , which is 1s.  $+\frac{82}{203}$  i.e. 1.4s. It will be noted, first, that in calculating the median, none of the difficulties are encountered in this case which were met with when the average was being estimated. Secondly, the median and average are not the same ; the average was 1.1s

We have so far considered, in our processes of analysing tabulated data the calculation of averages for purposes of making comparisons. These averages are used as simple representatives of cumbersome data ; by comparing averages we can get crude comparisons between the various pieces of information in our tables, but we realize, of course, that by resorting to averages for this purpose

we have sacrificed a good deal of information at our disposal. It is wise, therefore, whenever possible, to supplement the information conveyed by an average with further detail. For instance, in the table on p. 65 relating to average wagon loads, the final figures for Great Britain in July, 1929, and July, 1930, are 5.51 tons and 5.41 tons respectively. But actually, as these figures are representative of such widely different areas and classes of freight, it is thought right to divide the country into a number of districts and to divide the freight traffic into the three classes shown in that table. Thus, the final average figures are qualified by the other figures in the table, and we see that the  $5\frac{1}{2}$  tons average figure is representative of figures ranging from 2 tons to  $12\frac{1}{2}$  tons.

The table on p. 95, referring to income tax, again shows how little information is actually conveyed by an average; the incomes range from £2,000 to over £100,000 a year, the average is £5,750.

It is useful to us to have some idea of the amount of variation in the group of measurements which we are considering. No knowledge of this is conveyed by an average. If we say that the average wage of a group of men is 45s there is nothing in this statement to indicate whether this is an average of a group whose wages range from 40s. to 50s., or of a group whose wages range from 30s. to 70s. Without some indication of the manner in which the characteristic is distributed throughout the group, the average alone is a poor representative, and any comparisons made solely on averages will have the defect of crudeness.

The height of a person conveys to others a certain amount of information about his size, but, if we also know

his chest measurement and waist measurement, we are the more able to form a mental picture of this person than from a knowledge of height alone. We have now to consider other statistics of a group of measurements besides the average.

In practice, as far as possible, when a representative figure, such as an average, is chosen for a group, it is supplemented by other figures which convey an idea of the extent of variation in the group. So far we have considered the use of the average and the median as representatives. Just as there are these alternative possibilities in the choice of a representative, and as we see later there are other methods of representing a group, so there are alternatives in our choice of figures to indicate the amount of variation in the group.

One of the simplest methods is to use the actual range of variation in the group. Thus on p. 112 there are 236 districts with unemployment percentage ranging from 0.3 to 69.5, a total range of 69.2. There are disadvantages in using the range. In the first place we cannot, as a rule, obtain it from the usual table. For instance, we cannot get the range of variation in profit and loss in the Coal Mining Industry from the table on p. 91, because the lowest and highest figures are not specified in such a table. Secondly the range depends on two only of the values and, if one of the extreme values were very much different from its nearest neighbouring values, the inclusion of that value in the group might make a considerable difference to the range. If the highest value 69.5 were not in the group of Unemployment Percentages, the range would be altered to 61.9, which shows a great change from 69.2. At the same

time the general disposition of the other members of the group has remained unaltered. We want a figure which will give us an idea of the dispersal of the measurements in the group, and the range itself for these reasons is not considered generally suitable.

It is preferable to rely upon a figure which depends for its size on the general disposition of the measurements of the group, and not merely on the two extreme values. The general variation in the group can be indicated by means of the differences between the individual measurements and the representative figure. These differences are called *Deviations*.

If the average is the representative figure, the differences between the individual figures and the average are the deviations; some of these are necessarily positive and some are negative, the sum of all the deviations is zero. If the median is the representative figure, the deviations are again positive and negative, but the sum of them is not necessarily zero.

A useful measure of the extent of the variation in the group is the *average deviation* or *mean deviation*. This is simply obtained by considering the deviations (without signs) as a group of measurements and getting the average of them. Obviously, if there is large variation amongst the individual measurements, there will be large deviations from the representative figure, and the mean deviation will be large. If the measurements in the group are all very close to one another, they will also be close to their representative, the deviations will all be small and the mean deviation will also be small. Thus the mean deviation will serve as an indication of the extent of little or great variation in the whole group.

Moreover to its calculation all members in the group will contribute something.

There are two measurements of this kind: the mean deviation from the average and the mean deviation from the median. Thus if we had a small group of wages: 45s., 46s. 6d., 47s., 48s., 48s. 6d., 49s., 49s., 50s., of which the average is 47s. 10½d., and the median 48s. 3d., the deviations are:—

	<i>From the average</i>	<i>From the median</i>
	— 2/10½	— 3/3
	— 1/ 4½	— 1/9
	— /10½	— 1/3
	+ / 1½	— /3
	+ / 7½	+ /3
	+ 1/ 1½	+ /9
	+ 1/ 1½	+ /9
	+ 2/ 1½	+ 1/9
Sum of deviations (ignoring signs)	10/ 3	10/0
Mean deviation	1/3½	1/3

The calculation of the mean deviation, when the information regarding the group is given in the form of a table, again presents difficulties, just as did the calculation of the average. Let us consider as an illustration the data respecting the marks obtained by 133 candidates in an examination, shown on p. 84:—

Marks	35-	40-	45-	50-	55-	60-	65-	70-	75-	Total
Number of Candidates	1	5	12	34	23	22	12	1		133

The average was estimated at 58.2 marks.  
definitely the facts set out in the table below:—

We know  
22

<i>Groups with negative deviations from the average ranging between</i>	<i>Number of Candidates</i>
23.2 & 19.2	1
18.2 & 14.2	5
13.2 & 9.2	12
8.2 & 4.2	34
 <i>Group with both positive &amp; negative deviations .</i>	 23
 <i>Groups with positive deviations from the average ranging between</i>	
1.8 & 5.8	22
6.8 & 10.8	23
11.8 & 15.8	12
16.8 & 20.8	1

We assume, as before, for purposes of obtaining an approximation to the mean deviation, that the average deviation in the constituent groups is half-way between the limits between which the deviations range. Thus, we assume that, of the 34 candidates with deviations ranging between 8.2 and 4.2, the average deviation is 6.2 and therefore obtain an estimate  $34 \times 6.2$  for the contribution of this group to the total deviation, from which we are to get the mean deviation. We further assume that the middle group of 23, containing individuals some of whom have negative deviations and some of whom have positive deviations from the average, can be split up into two, the one containing those individuals with marks ranging from the lower limit of the group to the middle and the other containing those with marks ranging from the middle to the upper limit of the group, and the mean deviations in these two groups can be estimated by the method of proportion, based on the assumption that the

whole 23 are evenly distributed between the limits of the group. Thus, since the range of marks in this group is 5 and the central mark is 57, we assume the existence of a group, with negative deviations, with marks from  $54\frac{1}{2}$  to 58.2, a range of 3.7, and a group, with positive deviations, with marks from 58.2 to  $59\frac{1}{2}$ , a range of 1.3. (It will be observed that we are taking the range of the whole group as from  $54\frac{1}{2}$  to  $59\frac{1}{2}$ ; this is consistent with the original marks given to the nearest whole number. A mark of 55 really stands for any mark between  $54\frac{1}{2}$  and  $55\frac{1}{2}$ ; if 300 were the maximum it is possible that three candidates, allotted 55 marks when the maximum is 100, might be awarded 165, 164, 166, corresponding to 55.0, 54.7, 55.3.) We split this group of 23 candidates into two parts, assuming that 17 have negative deviations from 0 to 3.7, and 6 have positive deviations from 0 to 1.3, (17 is  $\frac{3.7}{5} \times 23$  and 6 is  $\frac{1.3}{5} \times 23$ ). Then we do as before, assume that the 17 candidates have an average deviation of 1.85 (half-way between 0 and 3.7) and the 6 candidates have an average deviation of 0.65 half-way between 0 and 1.3). Thus for the calculation of the mean deviation we have the following table:—

	<i>Average deviation</i>	<i>Candidates</i>	<i>Total deviation</i>
Negative	21 2	1	21 2
	16 2	5	81.0
	11 2	12	134 4
	6.2	34	210 8
	1.85	17	31 45
Positive :	0.65	6	3 90
	3.8	22	83 6
	8.8	23	202.4
	13.8	12	165.6
	18.8	1	18 8
		<hr/> 133	<hr/> 953.15

Errors enter into the calculation of the total deviation due to the average mark of those in a small 5-mark group being assumed half-way along the range of variation in the group. These errors tend to cancel when the average is being calculated, but are cumulated when the mean deviation is being obtained. An approximate correction on this account has been calculated theoretically, and consists in subtracting from the total deviation so far obtained,  $\frac{1}{2} \times \text{range of the small groups} \times \text{the number in that middle group in which the individual marks range between limits which include the average}$ . This correction in our case is  $-\frac{1}{6} \times 5 \times 23 = -19.17$ . Thus the total deviation is  $953.15 - 19.17 = 933.98$ .

The mean deviation is  $\frac{933.98}{133} = 7.0$ .

The closeness of this approximation may be tested by reference to the correct mean deviation calculated from the original details shown on p. 87. The average is 58.0 (p. 88), and the deviations, with the number of candidates having these deviations from the average, are shown on p. 127. Thus the estimate made from the table is 0.4 too great.

Another method of measuring the amount of variation in a group of measurements is by means of the *Standard Deviation*. This measure is calculated as a representative of the individual deviations from the average in the following way. the deviations from the average are squared and these squares are averaged, the square root of this average is the standard deviation. Thus, if we had seven measurements, 8, 6, 9, 7, 7, 10, 9, of which the average is 8, the deviations from the average are 0, -2,

$+1, -1, -1, +2, +1$ , the squares of these, 0, 4, 1, 1, 1, 4, 1, are averaged,  $\frac{12}{7} = 1.714$ . The standard deviation

Negative Deviations	No of Candidates	Total Deviation
20	1	20
16	2	32
14	3	42
13	1	13
12	1	12
11	1	11
10	4	40
9	5	45
8	6	48
7	4	28
6	3	18
5	7	35
4	14	56
3	10	30
2	3	6
1	5	5
0	4	—
		441
1	1	1
2	6	12
3	6	18
4	2	8
5	7	35
6	1	6
7	9	63
8	6	48
9	4	36
10	4	40
12	6	72
13	3	39
14	1	14
15	2	30
19	1	19
		441
Total deviation		882
Mean deviation		$\frac{882}{133} = 6.6$

In order to save trouble in the calculations we adopt, at this stage, a simple device derived from the formula for the standard deviation. This latter is the square root of

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n},$$

which may be written as

$$\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n} - \frac{2\bar{x}(x_1 + x_2 + \dots + x_n)}{n} + \bar{x}^2,$$

but since  $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$ , the above reduces to

$$\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n} - \bar{x}^2.$$

Thus if we measure our  $x$ 's from any zero whatsoever and calculate  $\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}$ , then subtracting  $\bar{x}^2$ , will give us the desired quantity from which to extract the standard deviation. Instead, therefore, of calculating the sum of the quantities in the last column of the above table we calculate  $x_1^2 + x_2^2 + \dots + x_n^2$  from  $x$ 's measured from 57 as we did on p. 89.

Values of $x$	Number of Candidates	For total $x^2$
- 20	1	$1 \times 20^2 = 1 \times 5^2 \times 16 = 5^2 \times 16$
- 15	5	$5 \times 15^2 = 5 \times 5^2 \times 9 = 5^2 \times 45$
- 10	12	$12 \times 10^2 = 12 \times 5^2 \times 4 = 5^2 \times 48$
- 5	34	$34 \times 5^2 = 34 \times 5^2 \times 1 = 5^2 \times 34$
0	23	0
+ 5	22	$22 \times 5^2 = 22 \times 5^2 \times 1 = 5^2 \times 22$
+ 10	23	$23 \times 10^2 = 23 \times 5^2 \times 4 = 5^2 \times 92$
+ 15	12	$12 \times 15^2 = 12 \times 5^2 \times 9 = 5^2 \times 108$
+ 20	1	$1 \times 20^2 = 1 \times 5^2 \times 16 = 5^2 \times 16$
	133	$5^2 \times 381$

$$\bar{x} = 5 \times 0.233 \text{ (see p. 89)} \quad 381 - 133 = 2 \ 865$$

For the standard deviation we have then :—

$25 \times 2.865 - 25 \times 0.233^2$ , which is  $25 \times 2.865 - 25 \times .054$ , i.e.  $25 \times 2.811$ .

A sensible correction has now to be applied to counter-balance the assumption that the middles of the ranges of the groups in the table are identical with the average marks of these groups, just as a similar correction was necessary in the case of the mean deviation. This is Sheppard's correction, and it consists of subtracting  $\frac{1}{12}$ th of the square of the ranges of the groups; in our illustration we therefore subtract  $\frac{1}{12} \times 5^2$ . This gives  $25 \times 2.811 - 25 \times .083 = 25 \times 2.728$ .

The standard deviation is  $5\sqrt{2.728} = 5 \times 1.649 = 8.2$  marks.

We can compare this result with the standard deviation obtained from the actual deviations (see p. 127).

Negative Deviations	Number of Candidates	(Deviations) <sup>2</sup>	Positive Deviations	Number of Candidates	(Deviations) <sup>2</sup>
20	1	400	19	1	361
16	2	512	15	2	450
14	3	588	14	1	196
13	1	169	13	3	507
12	1	144	12	6	864
11	1	121			
10	4	400	10	4	400
9	5	405	9	4	324
8	6	384	8	6	384
7	4	196	7	9	441
6	3	108	6	1	36
5	7	175	5	7	175
4	14	224	4	2	32
3	10	90	3	6	54
2	3	12	2	6	24
1	5	5	1	1	1
0	4	0			

Total (deviation)<sup>2</sup> is 8182. The average (deviation)<sup>2</sup> is  $\frac{8182}{133} = 61.52$ .<sup>1</sup> The standard deviation is  $\sqrt{61.52} = 7.8$ .

The standard deviation calculated from the table was 8.2, 0.4 too great.<sup>1</sup>

### *Quantile Deviation*

A further method of measuring the amount of variation in the group is by means of the *Quantile Deviation*.

This measure is associated with the median, which is sometimes used as a representative of the group. The median is the measurement of the middle individual in the group when all are arranged in order. The quartile measurements are similarly those of the individuals *standing half-way between the extremes and the median*. The median and quartiles divide the whole group into four equal parts. The quartiles are referred to as Lower Quartile and Upper Quartile, the lower quartile being less than the median, which in turn is less than the upper quartile. If there are 103 individual measurements in a group arranged in order of size, starting with the smallest, the lower quartile is the measurement of the 26th, the median is the measurement of the 52nd, the upper quartile is that of the 78th. Sometimes difficulties are encountered in the location of the median and quartiles because there are no individuals in the half or quarter positions. Thus, if there were 100 units in the group, the median would be assumed to be half-way between the 50th and 51st measurements, the upper quartile would be taken as half-way between the 75th and 76th measurements, and the

<sup>1</sup> Although the mean deviation may be calculated when the deviations are taken from the average or the median, the standard deviation is only computed from deviations from the average

lower quartile half-way between the 25th and 26th. If there were 102 units in the whole group, the lower quartile would be the measurement of the 26th individual, the median is half-way between the 51st and 52nd, the upper quartile is the measurement of the 77th individual.

The Quartile Deviation is defined as half the difference between the upper and lower quartiles.

As an illustration we may refer to the table on p. 112, where there are 236 units in the group. The median is the measurement half-way between the 118th and 119th measurements, 7.85, the lower quartile is half-way between the 59th and 60th measurements, the 59th is 4.0, the 60th is 4.3, and we take the lower quartile as 4.15. The upper quartile is half-way between the 177th and 178th measurements which are 14.6 and 14.7, the upper quartile is 14.65. The quartile deviation is  $\frac{14.65 - 4.15}{2} = 5.25$ .

The position of the quartiles in the scale of measurements can be approximately located either from the cumulative diagram as in the case of the median, or by calculation from the cumulative table. Thus Diagram 9 shows that the lower quartile is 4.3 per cent and the upper quartile 15.2 per cent. The following extracts from the table on p. 114 show that the lower quartile, between the 59th and 60th measurements, is 3.95 +

Percentage under	Number of Districts	Percentage under	Number of Districts
3.95	57	9.95	140
5.95	94	14.95	178

$$\frac{59\frac{1}{2} - 57}{94 - 57} \times 2 = 3.95 + \frac{2\frac{1}{2}}{37} \times 2 = 3.95 + .13 = 4.1;$$

and the upper quartile, between the 177th and 178th measurements is  $9.95 + \frac{177\frac{1}{2} - 140}{178 - 140} \times 5 = 9.95 + \frac{37\frac{1}{2}}{38} \times 5 = 9.95 + 4.93 = 14.9$ .

We have then the following results:—

	From the Original Data	From the Cumulative Diagram	From the Cumulative Table
Lower Quartile . .	4.15	4.3	4.1
Upper Quartile . .	14.65	15.2	14.9
Quartile deviation . .	5.25	5.45	5.4

If we take as a further illustration the table of marks of the 133 candidates, and use the cumulative table, which is:—

Marks below	40	45	50	55	60	65	70	75	80
Number of Candidates	1	6	18	52	75	97	120	132	133

The median is the number of marks obtained by the 67th candidate, when they are arranged in order of merit, and is given by  $55 + \frac{67 - 52}{75 - 52} \times 5 = 55 + \frac{15}{23} \times 5 = 58.3$ .

The lower quartile is  $50 + \frac{33\frac{1}{2} - 18}{52 - 18} \times 5 = 50 + \frac{15\frac{1}{2}}{34} \times 5 = 52.3$ .

The upper quartile is  $65 + \frac{100\frac{1}{2} - 97}{120 - 97} \times 5 = 65 + \frac{3\frac{1}{2}}{23} \times 5 = 65.8$

The quartile deviation is 6.7 marks.

It is worth while collecting these results, in order to stress the difference between them.

	Marks		Marks
Average . . . . .	58 2	Median . . . . .	58 3
Mean deviation from the average . . . . .	7 0	Quartile deviation . . . . .	6 7
Standard deviation . . . . .	8 2		

It is necessary to emphasize the fact that these various constants are different. They are expected to be different. They are obtained after submitting the tabulated material to different processes, and different results are likely. It is not usual to work out all these measures in a particular case. If we desire to indicate by means of a single figure the extent of the variation in a group, we should find one only of these measures. The various measures which have been described are used on different occasions. Sometimes the tabulated material is in such a form that it is preferable to use the quartile deviation, indeed, this measure may be the only one of its kind which is capable of calculation. With other material it may be preferable to obtain the standard deviation, when its calculation is possible. But there is a danger into which the unwary may be led, due to this embarrassing choice of measures of variation. Naturally we only obtain a measure of variation so that we can readily compare the group from this point of view with another group. We should always see that, when making comparisons, we are comparing the same kind of measures with one another. It would be

wrong to compare the standard deviation of one group with the quartile deviation of another. The quartile deviation and standard deviation of the same group differ sensibly and such a comparison as that suggested would be vitiated from the start. It is just as wrong to do this as it would be wrong, in comparing two persons as to size, if the height were taken as the measure of size in one case, and the chest measurement were used as the measure of size in the other case.

## CHAPTER 9

### WEIGHTED SUMS AND WEIGHTED AVERAGES

So far in our discussion of the contributions to the total by the individual members of a whole group we have confined ourselves to the cases where each member is considered to have the same importance as the others. Now we have to deal with those cases where this is not so. There are times when we wish to obtain a total for a group of units, when these units may be considered as equivalent from some points of view but are not so from other points of view. For instance we may be concerned with a group of 200 persons travelling on a particular train. If we think of these as 200 "souls" each unit is equivalent to all the others; if we think of them from the point of view of the Railway Company we may wish to obtain an idea of the seating capacity of the train required to carry them, and neglect the "infants in arms" and perhaps arrive at a total of 180 persons, or we may have to think of them as first class ticket holders, or third class ticket holders, or as persons travelling at half rates (e.g. children) or as persons travelling at special rates (e.g. workmen's ticket holders or cheap fare ticket holders). Thus a whole group may be considered homogeneous in one sense, and at the same time heterogeneous in other senses, and we are led to the notion of expressing certain of the members of the group in terms of "equivalent units", in order to get a total for the group which can be considered properly comparable with some other similar total.

We may consider an example of this kind of treatment supplied by the population of England and Wales. In 1881 the total population was 25,974,000, in 1921 it was 37,887,000, an increase of 46 per cent. Dr. E. C. Snow (*Journal of the Royal Statistical Society*, 1929, part iii, p. 333) is considering the members of the population as consumers of commodities, and realizing that the consumption-demand varies somewhat with age, suggests a scale, representing the average equivalent consumption-demand for persons in different age-groups, taking unity as the maximum consumption-demand (for persons aged 30). The scale is shown below:—

Males or Females, Age	0-14	15-29	30-44	45-59	60-74	75-90
Equivalent Consumption-demand	.19	.81	.95	.68	.32	.06

Thus he suggests that the average consumption-demand of children under 15 is .19, that of persons aged 30 being taken as unity. If we divide the population into age-groups and apply these figures as multipliers we shall obtain totals, for the two years, which may be compared. The details are shown top of p. 138.

Thus the population at these two dates consist of 16,300,000 and 22,600,000 consumption-demand units, i.e. the two populations are equivalent to these numbers of persons aged 30. The increase from 1881 to 1921 is 39 per cent, which differs considerably from the 46 per cent change in the actual numbers.

Again, in the London Survey, already referred to p. 9, the following scale is used to indicate housing needs of

Age	Males and Females		Equivalent Con- sumption- demand	Consumption- demand Units	
	1881 (000)	1921 (000)		1881 (000)	1921 (000)
0-14	9 469	10 500	19	1,800	1 995
15-29	6 923	9 615	81	5,600	7,790
30-44	6 686	8 148	95	6,350	7,740
45-59	2 981	6 051	68	2 020	4,110
60-74	1 580	2 925	32	505	935
75-90	336	648	06	20	40
	25,974	37,887		16,295	22 610

different persons. Men aged 18 and over and women aged 16 and over are taken as adult units, boys 14 to 17, and girls of 14 and 15 are considered as  $\frac{3}{4}$  of these units, children of both sexes from 5 to 13 are taken as  $\frac{1}{2}$  units, and infants 0 to 4 as  $\frac{1}{4}$  units. We can get the total number of equivalent adults from the table below —

	1881 mns	1921 mns		Equivalent Adults 1881	Equivalent Adults 1921
Male and Female 0-4	3 5	3 3	$\frac{1}{4}$	0 88	0 82
5-13	5 4	6 4	$\frac{3}{4}$	2 7	3 2
Males " 14-17	1 0	1 4	$\frac{3}{4}$	75	1 05
Females 14-15	0 5	0 7	$\frac{3}{4}$	38	53
Males 18 and over and Females 16 and over	15 5	26 0	1	15 5	26 0
	26 0	37 9		20 2	31 6

Thus the comparative number of equivalent adults, from the point of view of housing needs, has changed from 20·2 millions to 31·6 millions, an increase of 51 per cent.

The number of "persons" has increased by 46 per cent between these two dates, considered as consumers

the increase is only 39 per cent, considered from the point of view of housing needs the increase is 51 per cent.

Two further illustrations of the necessity for the consideration of the contributions of the different members of a group being regarded as of different relative importance may be quoted.

When the total output of Coal in Germany is being considered for the purpose of making comparisons, Lignite is either separated from Bituminous Coal or included with it by expressing it in terms of Bituminous. Thus, in the *Report of the Royal Commission on the Coal Industry* (1925), p. 243, we read: "The additional 64 million tons of Lignite raised in 1925 represents, in terms of Bituminous Coal, at least 15 million tons, or substantially more than the reduction in consumption of Bituminous Coal in the country."

A further illustration is taken from the Special Memorandum No 8, *The Physical Volume of Production*, by J. W. F. Rowe, of the London and Cambridge Economic Service. When discussing the question of preparing an Index of Production for the Paper, Printing, and Allied Trades, using imports of raw materials, the author says: "The different imported raw materials do not however all result in the same paper equivalent. The British Paper Makers' Association have suggested the following method of estimating the total paper equivalent. Deduct one half the weight of net wet pulp imports (both chemical and mechanical), and one-tenth the weight of dry mechanical. This gives absolutely dry pulp equivalent, to which should be added dry chemical imports. Absolutely dry pulp yields 90 per cent of its weight as paper. Esparto and other fibres also yield approximately 90 per cent of paper."

The following table shows how this method of obtaining the total equivalent weight of paper has been applied to statistics of imports for the first six months of 1928.

The monthly trade returns June, 1928, give the original figures.

Net Imports 1928 (Jan-June)		Scale of Dry Pulp Equi- valent	Dry Pulp Equi- valent	Scale of Paper Equi- valent	Paper Equi- valent
<b>Pulp :—</b>					
Chemical Dry (tons)	205,797	1.0	205,797		
Chemical Wet (tons)	9,136	.5	4,568		
Mechanical Dry (tons)	1,153	.9	1,038		
Mechanical Wet (tons)	316,042	.5	158,021		
<b>Total Pulp</b>	<b>532,128</b>		<b>369,424</b>	.9	<b>332,482</b>
<b>Esparto and other Fibres (tons)</b>	<b>156,160</b>			.9	<b>140,544</b>
					<b>473,026</b>

Thus the total net imports are equivalent to 473,000 tons of paper. The total tonnage of Paper Making Materials may be interesting, from the point of view of the transport of these materials from one place to another, but from the point of view of the paper trade it is necessary to consider this other total of paper equivalent.

These examples sufficiently illustrate the idea that, on occasion, the constituent parts of a total drawn from the different members must be considered not of the same relative degree of importance, but as definitely different, and that a scale indicating this must be introduced before the necessary total can be arrived at. Such aggregates

are referred to as weighted sums, the relative importance of the different items is indicated by "weights".

Generally, if the items are  $I_1, I_2, I_3, \dots, I_n$ , and the weights are  $W_1, W_2, W_3, \dots, W_n$ , corresponding to these items, the weighted sum is  $I_1W_1 + I_2W_2 + \dots + I_nW_n$ . In the above illustrations the various multipliers which have been used are called "Weights".

Similarly when we are averaging a series of items of different degrees of importance we use a "weighted average", which is obtained by relating the weighted sum to the total of the weights. Thus the weighted average is

$$\frac{I_1W_1 + I_2W_2 + \dots + I_nW_n}{W_1 + W_2 + \dots + W_n}$$

It will be observed that the ordinary simple average is a special case of the weighted average, when the weights are all equal to unity. In this case the weighted average

becomes  $\frac{I_1 + I_2 + \dots + I_n}{n}$ ,  $n$  being the number of items.

A weighted average is unaltered if all the weights are multiplied or divided by the same quantity. Thus if the weights are  $2W_1, 2W_2, \dots, 2W_n$ , instead of  $W_1, W_2, \dots, W_n$ , the weighted average,

$$\frac{2I_1W_1 + 2I_2W_2 + \dots + 2I_nW_n}{2W_1 + 2W_2 + \dots + 2W_n}$$

is the same as before. Thus the *absolute* size of the weights does not matter, the *relative* weights only are of importance. This fact is often of great value when a large number of calculations are involved, especially as it can also be shown that slight changes in the weights do not have any material influence on the weighted average, so long as the essential relative sizes of the weights are maintained. It means, in practice, that if the weights involved are rather large numbers they can be reduced to more manageable

proportions without affecting the resulting average. This is illustrated in the example below:—

Items	Weights (1)	Products	Weights (2)	Products
19	48	912	5	95
25	58	1450	6	150
23	28	644	3	69
26	54	1404	5	130
22	76	1672	8	176
24	49	1176	5	120
27	31	837	3	81
26	42	1092	4	104
21	19	399	2	42
22	23	506	2	44
	428	10092	43	1011
Weighted average (1) $\frac{10092}{428}$ (2) $\frac{1011}{43}$				
= 23.6 = 23 5				

The difference between these may not be of any significance in a particular case, and the gain due to the reduced computations may be quite considerable. This principle also operates in those cases where we may feel sure that the items to be averaged should be accorded weights indicating their relative importance, but where at the same time we may have difficulties in determining what these weights actually are. In such cases we may be able to decide on *approximate* weights, and so long as these approximations do give a fair idea of the relative importance of the items in the group, we can trust the resulting average, on the grounds that any slight inaccuracies in the weights will not materially affect the result of the averaging process. Such cases arise quite frequently.

When the items to be averaged are all very nearly the same size, e.g. percentage figures near 100, we can often reduce the computations by expressing each figure as an excess or defect of some standard figure, as in the following illustration :—

Items	Items from 100	Weights	Products
95	- 5	5	- 25
103	+ 3	15	+ 45
102	+ 2	13	+ 26
98	- 2	8	- 16
97	- 3	4	- 12
100	0	6	0
106	+ 6	9	+ 54
93	- 7	11	- 77
96	- 2	8	- 16
99	- 1	6	- 6
101	+ 1	7	+ 7
102	+ 2	12	+ 24
97	- 3	10	- 30
		114	- 26

$$\text{Weighted average from 100} = - \frac{26}{114} = - .2$$

Weighted average is 99.8

In such cases the saving in arithmetic is considerable, with, of course, a reduced chance of making computation errors.

It is interesting to consider, in the general case, how great is the difference between the ordinary simple average and the weighted average. We may say, generally, that if the allocation of weights to the items is such that the larger items have the smaller weights, and the smaller items have the larger weights, the smaller items have more influence in the formation of the weighted average,

and this is less than the simple average. On the other hand, if the larger weights are attached to the larger items, and the smaller weights to the smaller items, the opposite is the case, and the weighted average is greater than the simple average. But if the allocation of weights to the items is not apparently connected with the size of the items, that is, if larger weights are as much attached to larger as to smaller items, and similarly with smaller weights, the difference between the two kinds of average may be *inconsiderable*. We can show this algebraically, as below.

We are given  $n$  items,  $I_1, I_2, \dots, I_n$ , having weights  $W_1, W_2, \dots, W_n$ . Let us call the simple average of the items  $I$ , and the simple average of the weights  $W$ , then  $I = \frac{I_1 + I_2 + \dots + I_n}{n}$ ,  $W = \frac{W_1 + W_2 + \dots + W_n}{n}$ .

Let us obtain the deviations of the items and the weights from their averages, and call them  $i_1, i_2, \dots, i_n$ ; and  $w_1, w_2, \dots, w_n$ . Then  $I_1 = I + i_1$ ,  $I_2 = I + i_2$ ,  $\dots$  and  $W_1 = W + w_1$ ,  $W_2 = W + w_2$ ,  $\dots$ . We have obviously  $\frac{i_1 + i_2 + \dots + i_n}{n} = 0$ ,  $\frac{w_1 + w_2 + \dots + w_n}{n} = 0$ . The

process of obtaining the weighted average is set out below :—

Items	Weights	Products
$I + i_1$	$W + w_1$	$IW + W i_1 + I w_1 + i_1 w_1$
$I + i_2$	$W + w_2$	$IW + W i_2 + I w_2 + i_2 w_2$
$\vdots$	$\vdots$	$\vdots$
$I + i_n$	$W + w_n$	$IW + W i_n + I w_n + i_n w_n$
	$nW$	$nIW + i_1 w_1 + i_2 w_2 + \dots + i_n w_n$

Thus the weighted average is  $\frac{nIW + i_1w_1 + i_2w_2 + \dots + i_nw_n}{nW}$   
 $= I + \frac{i_1w_1 + i_2w_2 + \dots + i_nw_n}{nW}.$

The difference between the weighted average and the simple average is therefore  $\frac{i_1w_1 + i_2w_2 + \dots + i_nw_n}{nW}.$

Now, some of the  $i$ 's are positive and some are negative, so also with the  $w$ 's; and if positive  $i$ 's tend to be associated with positive  $w$ 's, and negative  $i$ 's with negative  $w$ 's, then this difference will be positive and the weighted average will be greater than the simple average. Here is the case of the larger items having the larger weights and the smaller items the smaller weights. On the other hand, if the positive  $w$ 's tend to be associated with negative  $i$ 's and negative  $w$ 's with positive  $i$ 's, the sum of the products ( $iw$ ) will be negative and the weighted average will be less than the simple average. This is the case of the larger items having the smaller weights and the smaller items the larger weights. But if we find that there is no particular allocation of the weights to the items, determined by size, then, sometimes positive  $w$ 's will be found with negative  $i$ 's as well as with positive  $i$ 's; and negative  $w$ 's will also be associated indiscriminately with positive and negative  $i$ 's. In this case, the products ( $iw$ ) will be irregularly positive and negative and a good deal of cancellation occurs when the total  $i_1w_1 + i_2w_2 + \dots + i_nw_n$  is obtained, so that this total may be quite small, and the difference between the weighted average and the simple average will be smaller, since this total is divided by the sum of the weights,  $nW$ . In this case the two kinds of average may be, for practical purposes, the same.

Let us take a simple illustration :—

(1) Large items with large weights, small items with small weights.

<i>Signs of Deviations from averages</i>					
<i>Items</i>	<i>Weights</i>	<i>Products.</i>	<i>Items</i>	<i>Weights</i>	<i>Products</i>
1	1	1	—	—	+
2	2	4	—	—	+
3	3	9	—	—	+
4	4	16	—	—	+
5	5	25	—	—	+
6	6	36	+	+	+
7	7	49	+	+	+
8	8	64	+	+	+
9	9	81	+	+	+
10	10	100	+	+	+
<u>55</u>	<u>55</u>	<u>385</u>			

$$\text{Simple average } \frac{55}{10} = 5.5$$

$$\text{Weighted average } \frac{385}{55} = 7.0$$

(2) Large items with small weights, small items with large weights.

<i>Signs of Deviations from averages</i>					
<i>Items.</i>	<i>Weights</i>	<i>Products</i>	<i>Items</i>	<i>Weights</i>	<i>Products</i>
1	10	10	—	+	—
2	9	18	—	+	—
3	8	24	—	+	—
4	7	28	—	+	—
5	6	30	—	+	—
6	5	30	+	—	—
7	4	28	+	—	—
8	3	24	+	—	—
9	2	18	+	—	—
10	1	10	+	—	—
<u>55</u>	<u>55</u>	<u>220</u>			

$$\text{Simple average} = 5.5$$

$$\text{Weighted average} = \frac{220}{55} = 4.0$$

## (3) Items and weights not associated.

<i>Signs of Deviations from averages</i>					
<i>Items</i>	<i>Weights</i>	<i>Products</i>	<i>Items</i>	<i>Weights</i>	<i>Products</i>
1	10	10	—	+	—
2	3	6	—	—	+
3	6	18	—	+	—
4	4	16	—	—	+
5	5	25	—	—	+
6	8	48	+	+	+
7	2	14	+	—	—
8	1	8	+	—	—
9	9	81	+	+	+
10	7	70	+	+	+
<hr/> 55	<hr/> 55	<hr/> 296	Simple average = 5.5		
			Weighted average = $\frac{296}{55} = 5.4$		

In the last, the weights were given to the items in a haphazard fashion determined by the code

W	E	I	G	H	T	D	A	V	R
10	3	6	4	5	8	2	1	9	7

This appreciation of the reason why the weighted average and the simple average differ is very profitable in certain cases, where it is known definitely that the average of a group should be a weighted average—as the items should be considered as having different degrees of importance—but where we have great difficulty in deciding the weights. We may have reason for believing that the weights, if known, would not be in the slightest degree influenced by the size of the items, that, as likely as not, large and small weights would be associated indiscriminately with large and small items. If this is the case we are justified in taking the simple average as a good approximation to the unknown weighted average.

There are occasions when a weighted average of a series

of items is really an ordinary simple average. For instance, the male population of England and Wales in 1921 was 18,075,000, and the average number of deaths of males in 1920, 1921, 1922 was 240,605, the ratio  $\frac{240,605}{18,075,000} \times 1,000 =$

13.3 is called the crude death-rate. Now the male population may be divided into age-groups, and the numbers dead may also be so divided, and the death-rates of males of different ages may be calculated, to indicate the incidence of death at different ages.

ENGLAND AND WALES (MALES)

Ages	(1) Population 1921 (000)	(2) Deaths (Average 1920-1922)	(3) Death-rate (per 1,000)
Under 5	1,681	55,361	32.9
5-	1,767	5,151	2.9
10-	1,837	3,314	1.8
15-	1,728	4,901	2.8
20-	1,488	5,447	3.8
25-	2,621	11,551	4.4
35-	2,496	17,004	6.8
45-	2,133	25,073	11.8
55-	1,383	34,639	25.0
65-	730	42,025	57.6
75-	225	29,685	131.6
85-	25	6,455	259.9
	18,075	240,605	13.3

This figure, 13.3, may also be considered as the average of the death-rates in the different age-groups; if it is so regarded it is as a weighted average, the weights being the numbers in these age-groups. In the table above column (3) is obtained by dividing the figures of column (2) by those in column (1). If we consider the figures in

column (3) as a series of items to be averaged, the weights being the figures in column (1), the figures in column (2) will be the products of the items and weights, and the weighted average is obtained by dividing the total of column (2) by that of column (1).

Generally, a simple average may be calculated from totals relating to a whole group, and at the same time similar simple averages may be calculated from corresponding figures belonging to the various constituent parts into which the whole group has been divided. The simple average for the whole group will be a weighted average of the similar averages of the constituent parts. As a further illustration, the rate of growth of the population of England and Wales from 1921 to 1931 may be expressed as 5·4 per cent of the population in 1921; the rates of growth of the different counties may also be calculated. The figure 5·4 for the whole country will be the weighted average of these rates of growth, calculated for the counties, the weights being the county populations in 1921.

On p. 72 certain average outputs per man-shift were worked out. It will now be realized that these averages can be regarded either as weighted averages or simple averages. We showed that if in the six districts the following number of man-shifts in the standard period were worked.—

District	I	II	III	IV	V	VI	Total
Man-shifts (000)	36,010	15,045	42,145	58,149	84,121	35,637	271,107

Then the total output would have been 4,866·4 million cwt. in 1924, first quarter, giving an average output per

man-shift of 17.95 cwt. But we may now think of this last figure as the weighted average of the output per man-shift for each district, the weights being the number of man-shifts in the standard period. The calculations shown on p. 72 are merely those performed when working out the weighted average. As a matter of interest we may rework these weighted averages, taking as weights 36, 15, 42, 58, 84, 36 instead of the original figures. We should get the results detailed below:—

Weights	Items (1)	From 18 00	For Weighted Average	Items (2)	From 18 00	For Weighted Average
36	18 95	+ 95	+ 34 20	19 01	+ 1 01	+ 36 36
15	17 12	— 88	— 13 20	18 05	+ 05	+ .75
42	17 15	— 85	— 35 70	17 91	— 09	— 3 78
58	16 19	— 1 81	— 104 98	16 36	— 1 64	— 95 12
84	20 60	+ 2 60	+ 218 40	20 30	+ 2 30	+ 193 20
36	14 88	— 3 12	— 112 32	14 79	— 3 21	— 115 66
271			— 13 60			+ 15 85
Weighted average (1)			Weighted average (2)			
$18\ 00 - \frac{13\ 60}{271}$			$18\ 00 + \frac{15\ 85}{271}$			
= 17 95			= 18 06			

These results are the same as those obtained previously using the original figures as weights.

## CHAPTER 10

### INDEX NUMBERS

Weighted averages are frequently used in the calculation of index numbers. We may be concerned with the measurement of the change from one period to another in a certain factor, but this change may not be susceptible of direct measurement, owing to the factor not being liable to numerical appreciation at any time. But evidence of this change is observed when measurements are made of quantities which are influenced by it. We can then collect statistics relating to a group of quantities, from which can be calculated the changes for each quantity in the measurements obtained. If these are expressed as percentages we have a group of them, each of which owes its size partly to the hidden factor about which we desire information. We can, by taking an average of the group changes, get an approximate notion of the change in this factor. We deal, in practice, with relative changes, as suggested above, and not absolute changes, because the quantities about which we have information may be measured in different units, and the final result is expressed in relative form, as a percentage, for this reason.

We may, for instance, be concerned with the general change in industrial productivity in a country over a period of time. Evidence of this change may be seen in a variety of ways, in changes in output of coal, of steel, of cotton yarns, of motor cars, of units of electricity, of tobacco manufactures, of beer, and so on. Some of the units involved in these will be tons, numbers, barrels, etc. We can reduce these changes to a common unit by

expressing them as percentages, or, as is most often done, instead of relating the *changes* in output of a particular industry or commodity to the output in the earlier period, we relate the *total* output in the one period to that in the other. In the latter case we get for our different industries or commodities a group of percentages, which reflect in some way or another the change which has taken place generally in industrial productivity. From this group we obtain an average which we take to indicate the level of industrial productivity at the one period compared with that at the other period. This average, which is a percentage figure, we call an Index Number of Industrial Production.

Similarly, we may be concerned with attempting to measure, generally, changes in the Price Level. Evidence of this change may be observed in changes in the prices of wheat, pig-iron, rubber, leather, raw cotton etc., and as these prices all involve different units, we express them as percentages, or, instead of concentrating on actual changes, we may express each price in the later period as a percentage of the corresponding price in the earlier period. These percentages form a group, the average is taken to indicate the general change from the earlier period to the later period, and is called for this reason a Price Index Number.

Generally, we may indicate the procedure in this way :—

Unit.	Statistics for First Period	Statistics for Second Period.	Percentages	Percentage Increase.
A	$a_1$	$a_2$	$a_2/a_1 \times 100$	$\frac{a_2 - a_1}{a_1} \times 100$
B	$b_1$	$b_2$	$b_2/b_1 \times 100$	$\frac{b_2 - b_1}{b_1} \times 100$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

The Index Number is that obtained as an average of the first column of percentages, but sometimes the average of the second column is used, and thus indicates the percentage *change* from the earlier period to the later.

As an illustration we may quote from the *Ministry of Labour Gazette*, January, 1929, where details are given of the calculation of the Cost of Living Index Number.

Article	Average Price		1st Jan 1929 compared with July 1914	
	July, 1914	1st Jan. 1929	Per-centage	Per-centage Increase
	<i>s d</i>	<i>s d</i>		
Beef - British				
Ribs . . . . . per lb	10	1 4½	168	68
Thin Flank . . . . . "	6½	9½	138	38
Beef Chilled or Frozen				
Ribs . . . . . per lb	7½	10½	143	43
Thin Flank . . . . . "	4½	5½	113	13
Mutton - British				
Legs . . . . . "	10½	1 6	174	74
Breast . . . . . "	6½	10	154	54
Mutton - Frozen				
Legs . . . . . "	6½	11½	169	69
Breast . . . . . "	4	5	127	27
Bacon (Streaky)	11½	1 4	143	43
Flour . . . . . per 7 lb	10½	1 3½	146	46
Bread . . . . . per 4 lb	5½	8½	149	49
Tea . . . . . per lb	1 6½	2 4½	155	55
Sugar (Granulated)	2	3	149	49
Milk . . . . . per qt	3½	6½	189	89
Butter:				
Fresh . . . . . per lb	1 2½	2 1	172	72
Salt . . . . . "	1 2½	1 11½	166	66
Cheese . . . . . "	8½	1 3	172	72
Margarine . . . . . "	7	7½	106	6
Eggs (Fresh) . . . . . each	1½	2½	231	131
Potatoes . . . . . per 7 lb	4½	6½	136	36
Fish . . . . . "	—	—	218	118
Average . . . . .			159	59

The *Ministry of Labour Gazette* generally quotes the increase in cost of living as a percentage, here 59 per cent. The index number is 159. This figure indicates the relative movement from the general price level of July, 1914, to that at the beginning of 1929.

The problem of obtaining an index number, thus outlined, appears to be a simple one. In practice, however, difficulties occur, first, in deciding upon a base period, second, in deciding what amount of information is to be included in the scope of the calculations, third, in deciding what weights shall be given to the items to be averaged, fourth, in deciding what kind of averaging process is to be used in the calculation of the index number.

With regard to the first point, it is convenient in many cases to relate later period figures to the same earlier period figures, i.e. to have a fixed base to which reference is made. Thus, in the Board of Trade Index of Production the year 1924 is the base year, the production in later years is expressed as a percentage of that in 1924. In choosing a base period consideration must be given to the question how far the statistics of this period may be regarded as reasonably normal. It is not convenient to choose as a period to which reference is continually being made one which is known to have been unusual, in the sense that it was a period, for instance, in which there was a great deal of labour unrest, or financial chaos, or a period of war. For if we did make such a choice, we should always have to qualify any index number which was calculated with some statement calling attention to the abnormality of the base period. For instance, after the War many comparisons were made with the immediate pre-War years, but, each time figures indicating these

comparisons were quoted, attention was drawn to the fact that in the year before the War conditions of trade, employment, and so on were unusually good—we were at a boom period. In order to evade this difficulty, instead of taking a single period as base for the purpose of computing an index number, an artificial base period is often chosen. The experience of a group of years is averaged and these averages are taken as the base figures to which subsequent figures are referred. Thus, in the original index number of production worked out for the London and Cambridge Economic Service by J. W. F. Rowe (see Special Memorandum No 8) the base period was 1907–1913. The output of coal, for instance, in a later year, say 1925, was expressed as a percentage of the average output in these years 1907–1913, and so for the other commodities whose output was included in the computations of the index number. The index was obtained as an average of these percentages. The wholesale price index number calculated by the *Statist* is referred to the average experience of the group of years 1867–1877, this period is the base period of the index number.

Since the War, for many index numbers the year 1924 has been chosen as base period, as it was considered to be the first post-War year which could be regarded as free from abnormal events, previous years' statistics had been disturbed by strikes, troubles abroad, and so on.

So far we have been concerned with the question of a fixed base period to which all subsequent years should be referred. From some points of view, there are advantages in having a moving base. In this case, instead of each period's figures being related to the fixed base period's figures, each period's figures are related to the preceding

period's figures. Thus, if we were dealing with annual figures, we would express the 1928 figures as percentages of those for 1927, and get an index number on 1927 as base; then for 1929 we would express the 1929 figures as percentages of those for 1928 and get an index number for 1929 on 1928 as base; similarly we would get an index number for 1930 based on 1929; and so on. This method, called also the chain base method, automatically gives the short period changes and is of advantage when the year to year movements are of the more interest than the change over a longer period of time. We can, of course, obtain from these a new series of index numbers referred to a fixed base. For instance, if we have the following figures —

Index for 1928 referred to 1927 is	125
" " 1929 " " 1928 "	110
" " 1930 " " 1929 "	105

Then  $\frac{125}{100} \times 110$  will give an Index number for 1929 on 1927 as base  
 "  $\frac{125}{100} \times \frac{110}{100} \times 105$  " " " 1930 " 1927 "

and we should have these index numbers on the base 1927

1927	1928	1929	1930
100	125	137.5	144.4

There is another advantage of the moving or chain base method over the fixed base method, which is concerned with the available information from which the index number is to be calculated. In a long period of time the evidence, which we are using as indicative of the changing nature of the factor we are dealing with, may alter somewhat, and it may be difficult to get proper comparisons. For instance, if we are dealing with a price index

number based on a period before 1900 (say), we cannot compare prices of motor cars in 1932 with similar prices in the base period. Prices current in the base period similarly may have no counterpart in the later period. The same kind of difficulty arises when the same description is used at two periods to cover substantially different articles, e.g. stockings made of wool at one period and stockings made of artificial silk at a later period. We may easily be able to make comparisons between successive years, but may find difficulty in establishing the proper contacts between years some time apart. In such a case the chain base method enables us to link up two distant years, and an attempt to make a direct comparison perhaps break down.

It is worth while, at this stage, examining the second of our considerations, viz. the information to be used in the calculation of the index number. The data used may be the whole information available or they may only be a sample of this. In certain cases sampling only is practicable, e.g. in the calculation of a cost of living index number. A glance at a grocer's shop window, or a list of prices of different "cuts" of meat, or a study of a store's list, is enough to make us realize what a large variety of articles may enter into consideration if we are dealing with changes in price of foods alone. Similar concentration on clothing again emphasizes the enormously great number of prices which we might consider. When we look carefully into the problem, we see that there is a comparatively small number of items to which we can restrict ourselves, which will account for the greater part of the ordinary necessary purchases in everyday life. In order to avoid this wide survey, the Ministry of Labour

in their calculations of the cost of living index number actually do restrict themselves to a sample of all the possible information, knowing that the sample chosen represents to a large degree all ordinary purchases. Thus the typical purchases include beef but not salt, tea and sugar, potatoes but not other vegetables, bread but not cake. It must be remembered also that the information sought does not consist of a list of articles purchased, but a list of prices of such articles. There are thousands of shops in the country which quote these prices; and, again, instead of attempting to survey the whole country, a sample is taken, first of all by choosing a number of towns and then by choosing a number of shops in these towns. The cost of living index number inquiry is essentially an inquiry by sample.

Similarly, in the construction of the *Statist* and the Board of Trade Wholesale Price Indices, a sample only of all the possible information is taken; in the former case forty-five items are included, in the latter case 150 items.<sup>1</sup> So in the case of the Index of Production issued by the London and Cambridge Economic Service. In this case, however, the problem is not one of sampling existing information but of taking account of as much information as is available. We can use data referring to output of coal because these are known, we cannot use similar figures for production of suits of clothes because they are not available. We know the output of pig-iron, but we do not know the output of bedsteads.

On the other hand, the indices published by the Board of Trade showing relative changes in value and volume of

<sup>1</sup> Actually, there are somewhat more than these, as some of the "items" are already averages of a few quotations.

the import and export trade do take account, as far as is possible, of all the available information.

Where a sample is used in the calculation of an index number, the sample is expected to be so representative, that the final index arrived at will reasonably show the changes with time in the factor which is being measured.

We have then a certain number of items in the form of percentages, which are to be averaged, and we come to the third point, the question of weights to be attached to these items. Palpably, weighting is necessary<sup>1</sup> because the items compose a heterogeneous group. If we are dealing with a retail food price index, for instance, and we have items 115, 125, 132, 116, 124, etc., which have been calculated from prices, at two dates, of beef, milk, bacon, bread, eggs, etc., we must realize the necessity for weights to indicate the relative degree of importance of these items, i.e. to indicate the relative degree of importance of these *articles*, beef, milk, bacon, bread, eggs, etc. The question is what considerations are to guide us in assessing the relative importance of these commodities? It is reasonable that we should assume that this should be decided by the consumers of them. The Ministry of Labour are guided in this connection by the tastes of working class households, and the importance of these articles is determined by the relative amounts spent on them in such households. If, for instance, the available money (say 10s.) is spent only on five articles and distributed in this way: Beef 3s., milk 2s., bacon 2s.,

<sup>1</sup> *A priori*, that is; unless we have reason to believe, on the principle discussed on p 147, that the weighted average and the unweighted average would not differ significantly

bread 2s., eggs 1s., then we may assume weights proportional to these, viz.: Beef 30, milk 20, bacon 20, bread 20, eggs 10, total 100. We determine the relative importance of the articles by the way the consumers spend their money on them. If three times as much is spent on beef as on eggs, then we assume that the relative importance of beef to eggs is three to one. The Ministry of Labour use as weights, in the cost of living index number, figures determined by many typical working class budgets. In these budgets the amounts spent on different articles of food are taken into account, also amounts spent on rent and rates, clothing, fuel and light, and so on, and the average experience is used as a guide to the relative importance of the different items included in the index number computations.

Similar considerations arise when we are dealing with a number of items expressing the relative output of certain industries or trades in a given period, compared with a base period. The average of these items is to be used as an index number of production. What method of weighting shall we adopt to indicate the relative importance of the different items, i.e. what figures will properly indicate the relative importance of different industries? Are we to use man-power, horse-power of machines used, value of product, or value of net output? Obviously, the medium of comparison is important. We must use weights which are proportional to figures measured in the same unit. The question is, what is the best unit for comparison between different industries? Actually, the unit which is used is one of value, and the weights taken are figures proportional to the net outputs of the various industries. The net output is the gross value of the product less the

cost of materials used. There are objections to using manpower and horse-power of machines. Mere numbers employed is no certain guide to the relative position of a particular industry in the national economy; if we are dealing with numbers we should have to qualify them with figures expressive of the relative powers and skill of the persons included, which would be difficult of achievement. Also, horse-power of machines is no guide, because some industries are more mechanically equipped than others, merely because of their particular nature, and out of all relation to the place these industries occupy in any scale of importance. Gross output, the total value of the product, is not a satisfactory guide, because the gross output of certain industries is high, since they are producing finished goods from nearly finished goods which are costly to those industries. The gross output of a firm manufacturing steel girders may be higher than that of a firm manufacturing pig-iron, but one would not for that reason assume that the production of pig-iron was not as important as that of steel girders. The net outputs of different industries are acknowledged to give the surest guide to their relative importance. In the two illustrations cited above the weights used in the calculation of the weighted average, which is the index number, are values. In many index number calculations the only possible medium of comparison of the units which form the heterogeneous group is that of value, measured in money. We can only compare tons of pig-iron, bushels of wheat, gallons of milk, thousands of barrels of beer, if we express each in terms of money value.

In certain cases no weights are used, e.g. in the *Statist* Index Number of Wholesale Prices and in the similar

index calculated by the Board of Trade. Here each item entering into the group is considered of the same degree of importance as the others, no attempt being made to assess them comparatively. The items which are averaged are relative prices between the later period and the base period of a number of raw materials and partly manufactured goods, such as raw cotton, pig-iron, coal, cotton yarns. It is hard to decide from what point of view we should attempt to determine the relative importance of these commodities. We saw above that the ordinary simple average would not differ greatly from the weighted average so long as the weights were unrelated to the sizes of the items. If this is the case here, then no appreciable error is introduced into the index number by taking this as an unweighted average.

Finally we have to consider what kind of averaging process is to be used in the calculation of the index number. We discussed earlier the use of the average and the median as representative figures of a group. The geometric mean of a group of numbers is also sometimes used as a representative. If the group consists of  $n$  values  $x_1, x_2, \dots, x_n$  the geometric mean is  $\sqrt[n]{x_1 x_2 \dots x_n}$ . It is usually computed by taking the simple average of the logarithms of the numbers. Thus the logarithm of the geometric mean is

$$\frac{1}{n} (\log x_1 + \log x_2 + \dots + \log x_n)$$

If the original data are given in the usual tabular form in grades the calculation of the geometric mean is impossible, but, when the original values of the item to be averaged are known, then little more computation is required than is necessary for the simple average. Just as a weighted

average can be calculated, so can a weighted geometric mean or a weighted median be obtained. The latter is not often used; to obtain the former, we get first the weighted average of the logarithms of the items. If the items are  $I_1, I_2, \dots, I_n$  and the weights are  $W_1, W_2, \dots, W_n$ , then the logarithm of the weighted geometric mean is

$$\frac{W_1 \log I_1 + W_2 \log I_2 + \dots + W_n \log I_n}{W_1 + W_2 + \dots + W_n}.$$

The geometric mean is of value when we are considering items in a group from the point of view of their relative differences rather than from the point of view of absolute differences, and is therefore reasonably used in index number computations, where the items averaged are themselves percentages. The geometric mean is also the more appropriate where the moving or chain base method is used, and certain theoretical considerations dealt with below support the claims of the geometric average. These considerations are concerned with the general problem of whether the index number performs satisfactorily its function of indicating the required change.

If we are dealing with a group of items, which show an average increase of (say) 25 per cent from one period to another, then the change can also be stated as a decrease of 20 per cent from the later period to the first period. The averaging process should show this. Suppose we take a simple illustration with two items.

	<i>Amount in</i>		<i>Year (2)</i>	<i>Year (1)</i>
	<i>Year (1)</i>	<i>Year (2)</i>	$\frac{\text{Year (2)}}{\text{Year (1)}} \times 100$	$\frac{\text{Year (1)}}{\text{Year (2)}} \times 100$
A	40	46	115.0	86.9
B	60	84	140.0	71.4
			<hr/>	<hr/>
Simple average			127.5	79.15
Geometric mean			126.9	78.8

Using the arithmetic average, we find that the year (2)'s figures are higher than those of year (1) by 27.5 per cent and therefore year (1)'s figures are lower than year (2)'s by 21.55 per cent of the latter. Actually from the index number we get 20.85 as this figure, ( $100 - 79.15$ ). The geometric mean shows year (2) as higher than year (1) by 26.9 per cent, which is equivalent to saying that year (1) is lower than year (2) by 21.2 per cent, agreeing with the index calculated for year (1) on year (2) as base.

( $\frac{100}{127.5} \times 100$  is not the same as 79.15, but  $\frac{100}{126.9} \times 100$  is equal to 78.8.) From the point of view of this test of the efficiency of an index number, the geometric mean is the more satisfactory, but of course in practice the difference between the results obtained by the two methods may be slight, and if the ordinary average method does not satisfy this test exactly, it may be held to do so to a sufficient degree of accuracy to warrant the method being used in the computation of these index numbers.

A geometric average is used in the calculation of the Board of Trade Wholesale Price Index Number, but a simple average is used in the calculation of that of the *Statist*.

It is interesting to compare the three following index numbers: The Board of Trade Wholesale Price Index Number, the *Statist* Wholesale Price Index Number, and the index number calculated by the Board of Trade showing the changes in average values of retained imports. The first two are different attempts to measure changes in wholesale prices, 150 items are used in the Board of Trade computations, forty-five items in those of the *Statist*. No weights are used, but the Board of Trade

uses a geometric mean and the *Statist* a simple average. The last is calculated from the average values of retained imports by means of a weighted average. As many of the available data as possible enter into the calculations and the weights are the values of the different descriptions of retained imports entering in 1924. Thus all these three are different in structure, yet as the following table shows, they tell nearly the same tale over a period of time.

Year	Wholesale Price Index Number		Board of Trade— Average value of Retained Imports
	Board of Trade	<i>Statist</i> <sup>1</sup>	
1924	100	100	100
1925	95.8	98	98.8
1926	89.1	90	90.4
1927	85.2	88	88.6
1928	84.4	98	87.7
1929	82.1	83	85.7
1930	71.9	69	75.6
1931	62.6	59	61.2

The agreement between these three series shows that approximately the same results are achieved, even though the sampling and the arithmetical processes differ.

There is one method of calculation of an index number which can be described in a somewhat different manner. It is that case where a weighted average is used for the calculation of a price index. Here the items are relative prices in the later period to those of the earlier period. The weights are the values of the articles for which we have price quotations, (or figures proportional to them), these values being expenditure on the articles by the

<sup>1</sup> The *Statist* figures have been reworked on a 1924 base

The other figures are taken from the *Journal of the Royal Statistical Society*, 1932, p. 614

average individual consumer, or the total amount expended by all consumers. We may set out the method of calculation as follows :—

Article	Price		Relative Price $\times 100$	Weights (Values)	Products
	Period (1)	Period (2)			
A	$p(a)_1$	$p(a)_2$	$p(a)_2/p(a)_1 \times 100 = I(a)$	$V(a)$	$I(a) \times V(a)$
B	$p(b)_1$	$p(b)_2$	$p(b)_2/p(b)_1 \times 100 = I(b)$	$V(b)$	$I(b) \times V(b)$
C	$p(c)_1$	$p(c)_2$	$p(c)_2/p(c)_1 \times 100 = I(c)$	$V(c)$	$I(c) \times V(c)$
D	$p(d)_1$	$p(d)_2$	$p(d)_2/p(d)_1 \times 100 = I(d)$	$V(d)$	$I(d) \times V(d)$
.	.	.	.	.	.

the index number is the weighted average :—

$$\frac{I(a)V(a) + I(b)V(b) + \dots}{V(a) + V(b) + \dots}$$

Now suppose that we have decided that the appropriate weights are the values in the first period (there may be a different distribution of values in the second period), these values are those of actual quantities consumed, and quantity  $\times$  price is equal to value. If we refer to the quantities as  $q(a)$ ,  $q(b)$ ,  $q(c)$ , . . . then we have  $q(a) \times p(a) = V(a)$ ,  $q(b) \times p(b) = V(b)$ , and so on, or for the period (1),  $q(a)_1 \times p(a)_1 = V(a)_1$ , and so on. Taking the first period values as weights, the index number may be written :—

$$\frac{\frac{p(a)_2}{p(a)_1} \times 100 \times V(a)_1 + \frac{p(b)_2}{p(b)_1} \times 100 \times V(b)_1 + \dots}{V(a)_1 + V(b)_1 + \dots}$$

which is equivalent to

$$\frac{p(a)_2 \times q(a)_1 + p(b)_2 \times q(b)_1 + \dots}{p(a)_1 \times q(a)_1 + p(b)_1 \times q(b)_1 + \dots} \times 100.$$

Now in the numerator,  $p(a)_2 \times q(a)_1$  may be considered as the amount of money which would be necessary to purchase the quantity of the article  $A$  at the later period's price,

Article	Quantities Purchased		Price		Expenditure	
			July, 1914	1st Jan. 1929	July, 1914	1st Jan. 1929
			s d	s d.	d	d.
Beef, British :						
Ribs . . .	1 lb	per lb	10	1 4½	10.0	16.8
Thin Flank . .	1 lb	"	6½	9½	6.5	9.3
Beef, Chilled or Frozen :						
Ribs . . .	1.35 lb.	"	7½	10½	9.8	13.9
Thin Flank . .	1.35 lb	"	4½	5½	8.4	7.4
Mutton, British :						
Legs . . .	0.48 lb	"	10½	1 8	5.0	8.8
Breast . . .	0.48 lb	"	8½	10	3.1	4.8
Mutton, Frozen :						
Legs . . .	0.75 lb	"	6½	11½	5.1	8.8
Breast . . .	0.75 lb	"	4	5	8.0	3.8
Bacon, Streaky .	1.1 lb	"	11½	1 4	12.4	17.6
Flour . . .	9 0 lb	per 7 lb	10½	1 3½	13.5	19.9
Bread . . .	23 5 lb	per 4 lb	5½	8½	33.8	49.9
Tea . . .	0.6 lb	per lb	1 6½	2 4½	14.6	22.8
Sugar, Granulated	6.1 lb.	"	2	3	12.2	18.3
Milk . . .	4.7 qts	per qt.	3½	6½	16.5	30.6
Butter, Fresh	0.95 lb	per lb	1 2½	2 1	13.8	23.7
" Salt . . .	0.98 lb	"	1 2½	1 11½	14.5	23.0
Cheese . . .	0.8 lb	"	8½	1 3	7.1	12.0
Margarine . . .	0.9 lb	"	7	7½	6.3	6.8
Eggs, Fresh	10	each	1½	2½	12.5	27.5
Potatoes . . .	17 lb	per 7 lb	4½	6½	11.6	15.8
Fish . . .	—	—	—	—	6.1	13.3
					223.8	354.6

$$\frac{354.6}{223.8} = 1.59$$

The Index number is 159

(The "Quantities purchased" figures in the above table are taken from *Prices and Wages in the United Kingdom, 1914-1920*, by Professor A. L. Bowley)

$p(b)_2 \times q(b)_1$  may similarly be considered as the amount necessary for the quantity of this article at its later period's price, and so with the other articles and their contribution to the numerator.

Thus the numerator may be reckoned as the total value of the first period's consumption revalued at the later period's prices. The denominator is, of course, the total value of the first period's consumption at prices then current. In practice we may have available the quantities consumed, and it may be a simpler procedure to revalue them at the later period's prices, rather than to work out the price ratios and compute the index number from the original formula. For instance, in the Ministry of Labour Retail Food Price Index, we can use the expenditure in a typical working-class budget on different articles, as weights attached to the relative prices, or we can revalue the quantities purchased at the later period's prices, and we shall get the same result either way. (Table on p. 167.)

This method is used by the Board of Trade in the calculation of the change in average values of retained imports, the figures quoted on p. 165. The quantities imported in the base year, (1924) are revalued at the average values of the later year (average value of a commodity is obtained by dividing the total value by the quantity). The total resulting from this revaluation is expressed as a percentage of the total value in the base year, giving the index number.

The price index number is sometimes expressed briefly in the form

$$\frac{\text{Sum} \left( \frac{p_2}{p_1} \times 100 \times V_1 \right)}{\text{Sum} (V_1)}, \text{ where the summation refers to}$$

the different articles or commodities, and in its new form as

$$\frac{\text{Sum } (p_2 q_1)}{\text{Sum } (p_1 q_1)} \times 100.$$

The Board of Trade also calculate an index number showing changes in volume of imports and exports. In this case the items to be averaged are  $q_2/q_1 \times 100$ , the relative quantities in the later period to those in the earlier period. In the averaging process the weights used are again the values of the different commodities imported or exported, and the index number is written briefly as

$$\frac{\text{Sum } \left( \frac{q_2}{q_1} \times 100 \times V_1 \right)}{\text{Sum } (V_1)}.$$

This may again be rewritten in the form

$$\frac{\text{Sum } (q_2 p_1)}{\text{Sum } (q_1 p_1)} \times 100.$$

Here the numerator consists of the sum of the results of revaluing the quantities of the later period at the average values of the earlier period. The denominator is the total value of the first period's quantities at the prices then current.

This method of stating the index number computation has the merit of putting it in a form which conveys something more definite to a person trying to appreciate the meaning of an index number, than is understood by a mere abstraction like a weighted average. But it must be remembered that the basic idea is that we are averaging certain percentages, and it may be considered as a fortunate "accident" that we are able to translate this into somewhat simpler terms in these cases mentioned. The

*Wholesale Price Index of the Board of Trade and the Statist* cannot be so treated.

## APPENDIX

### THE COST OF LIVING INDEX NUMBER

Reference is made to a pamphlet issued by the Ministry of Labour, "The Cost of Living Index Number : Method of Compilation."

"The statistics prepared by the Ministry of Labour are designed to measure the average increase in the cost of maintaining unchanged the pre-War standard of living of the working classes."

The foodstuffs included account for about 75 per cent of working class expenditure on food. Retail prices at the beginning of each month, of these, are obtained by the managers of Employment Exchanges from retailers in their localities engaged in a working-class trade. The information is collected in all towns with a population of 50,000 or more at the Census of 1911, and in 420 smaller towns and villages throughout the country. The total number of retailers is over 5,000.

The weights used are based on the average expenditure of 1,944 urban working class families, this information having been collected by the Board of Trade in 1904. The use of figures relating to 1904 instead of 1914 is considered reasonable on the grounds that between 1904 and 1914 no great change took place in the standard of living. The expenditure on margarine did change, and this item had special treatment. The weights given in the publication referred to above are :—

Beef . . .	48	Sugar . . .	19
Mutton . . .	24	Milk . . .	25
Bacon . . .	19	Butter . . .	41
Fish . . .	9	Cheese . . .	10
Flour . . .	20	Margarine . . .	10
Bread . . .	50	Eggs . . .	19
Tea . . .	22	Potatoes . . .	18
Total . . .	334		

The weighted average increase in the relative prices of foodstuffs is combined with similar figures indicating changes in rents, clothing, fuel, and light, and other items.

As regards rents, which includes rates and water rates, the information collected relates to that concerning controlled and decontrolled rents. Inquiries have been made respecting thirty-nine large towns on the subject of controlled rents, and from twenty-nine large towns about the proportion of working-class dwellings decontrolled, and the subsequent increase in rent.

As regards clothing, information is obtained as to changes in retail prices of men's suits and overcoats (ready-made and bespoke), woollen and cotton materials, underclothing and hosiery, and boots. Inquiry forms are completed each month by 300 outfitters, drapers, and boot retailers in eighty-one towns.

In the fuel and light group are included coal, gas, oil, candles, and matches. Information respecting prices of coal is obtained from twenty-nine principal towns, prices of gas are obtained from twenty-four towns, and similar details about lamp oil, candles and matches refer to forty-nine towns and ninety-one towns respectively. In the necessary aggregation of this information a weight of 6 is given to the relative increase in the price of coal, a weight of 3 to that of gas, and a weight of 0.7 is allocated to the change of price of oil, candles, and matches

taken together. These weights are determined by the relative expenditure on these items in the pre-War budgets already mentioned.

Amongst other items are soap and soda; domestic ironmongery, brushware and pottery, tobacco and cigarettes, fares; and newspapers. Prices of soap and soda are obtained from ninety-one towns, those of the next group are from forty towns. As to the rest the tobacco manufacturers retail price list, the principal transport undertakings, and the Ministry of Transport, and the daily Press supply the necessary information.

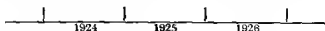
The combination of all this information into a single figure is performed by weighting. The weights used are food  $7\frac{1}{2}$ , rent 2, clothing  $1\frac{1}{2}$ , fuel and light 1, miscellaneous  $\frac{1}{2}$ , total  $12\frac{1}{2}$ . The budgets collected in 1904 showed that on the average about 60 per cent of the total expenditure was on food. As regards rents, a cost of living inquiry in 1912 showed that roughly one-sixth of working class expenditure was on rent, thus a weight of 16 out of a total of 100 is adopted. In the pre-War investigations the average expenditure on clothing was less than that on rent and, although there is wide variation from one household to another in this respect, a weight of 12 out of 100 is taken for this item. There are no extensive statistical data on the other items in the index number, but the available information suggests weights of 8 and 4 respectively out of a total of 100.

## CHAPTER II

### GRAPHS OF TIME SERIES

Diagrammatic representations of time series are often used to enable changes occurring to be easily appreciated. The normal procedure is to represent time along a horizontal scale and the quantity tabulated on a vertical scale. There are two kinds of time series which we have to consider. The first is a series of figures relating to some quantity which is measured at particular instants of time. Of this type are census of population figures, which give the population existing in a country on a given day in a particular year, or the estimates made annually by the Ministry of Labour of the total number of insured workers in July. The second type is a series of figures giving the aggregate experience in a number of time intervals, of some particular quantity. Of this type are the import and export figures giving totals for a series of months or years, or figures of total output of coal weekly.

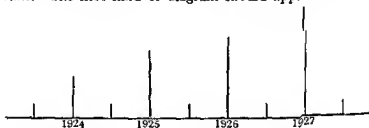
When we use a horizontal line in a diagram to represent time, we are quite rightly allowing lengths on the line to correspond to intervals in time. Thus if we take 1 inch to represent an interval of a year,  $\frac{1}{2}$  inch represents an interval of six months.



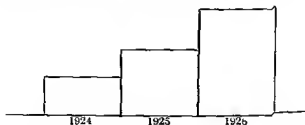
The mark on the scale separating one year from the next corresponds in time to midnight of New Year's Eve.

On such a diagram figures of the first kind referred to above should be shown by vertical ordinates erected at points on the horizontal scale corresponding to the particular instants of time to which these statistics refer. Thus annual figures of numbers of insured workers which are obtained for July each year, should be shown as vertical ordinates placed at points in the horizontal scale just over half-way between the points denoting the beginning and end of the year.

Those figures of the second kind should be shown by rectangular blocks erected on the space intervals corresponding to the different periods to which they refer. The first kind of diagram should appear thus:—

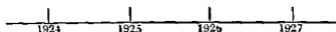


The second kind should appear in this form:—



The two kinds of diagram illustrate the different nature of the statistics which are being represented graphically.

Actually this method of procedure is not frequently adopted. The more usual course is to use, as representing a time interval in the horizontal scale, a point on the line instead of a space interval. Thus a series of points at equal intervals on the scale corresponds to a series of years, or months, or whatever time interval is involved.



Ordinates are erected at these points corresponding in length to the particular statistics which are being graphed, whether these figures relate to an instant of time or are aggregates over a period. Thus the same kind of graph would be used to show figures giving the mid-year estimates of population prepared by the Registrar-General, as for figures giving the output of coal during the year. The description accompanying the diagram would state clearly exactly what statistics are being graphically represented.

In this kind of diagram the space left between successive points on the horizontal scale is of a convenient size, arranged so that a person can read the diagram with ease. The points should not be too close together, otherwise the picture would show confusion, nor should they be too far apart, otherwise the diagram would be so large that the eye would undergo strain in attempting to absorb the details of a picture painted on a large canvas. The space between successive points naturally cannot correspond to a time interval since time intervals are being represented by points. On the other hand, if we had a diagram where (say) census populations were being shown, and

points on the horizontal scale corresponded to 1841, 1851, 1861, 1871, etc., then naturally points half-way between the original points shown would correspond to 1846, 1856, 1866, etc.

When ordinary graph paper is used, it is not necessary to draw the ordinates corresponding to the variable quantities

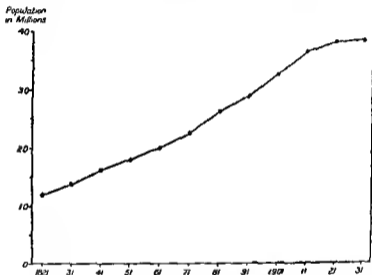


DIAGRAM 12 Population of England and Wales at each Census, 1821-1931.

which are being represented graphically, all that is necessary is to represent the end of the ordinate by means of a point on the paper. In order that a person reading the graph should be enabled easily to follow the movement in the values of the variable from one period to another, these consecutive points are usually connected by means of

straight lines. Usually there is no other significance to these straight lines than this, but sometimes we may be justified in inferring from the diagram an intermediate value of the variable for a period for which we had no data. Thus, suppose the census populations were plotted in this way, we might suppose that a point half-way along

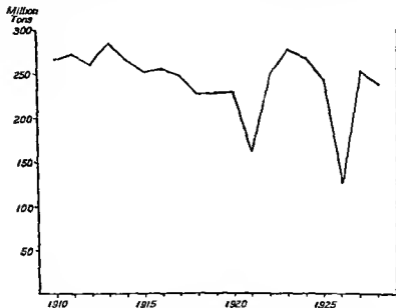


DIAGRAM 13 United Kingdom. Coal Production

the line connecting the tops of the ordinates corresponding to census populations of 1861 and 1871 would correspond to the population of 1866 on the assumption that the increase from 1861 to 1866 equalled that from 1866 to 1871.

Diagrams 12 and 13 illustrate this method of plotting graphs.

Sometimes the time interval is represented by a space interval along a horizontal scale and the ordinates, whether representing a variable which corresponds to an instant of time or to a period of time, are erected at a point in the middle of the space corresponding to the unit time interval.

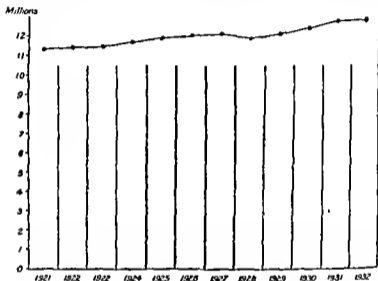


DIAGRAM 14 Number of Insured Workers (July each year)

Diagrams 14 and 15 illustrate this method of graphical representation.

It will readily be appreciated that, when we are plotting a graph showing large numbers, not much precision is possible. In Diagram 13, for instance, the points corresponding to the output of coal in 1919 and 1920 are at the same distance from the horizontal line, indicating that the output was the same in the two periods. Actually

these were 229,780,000 tons and 229,532,000 tons respectively. In such diagrams we must be satisfied that the gain, obtained by having a graph which enables us easily to trace the changes which take place, compensates

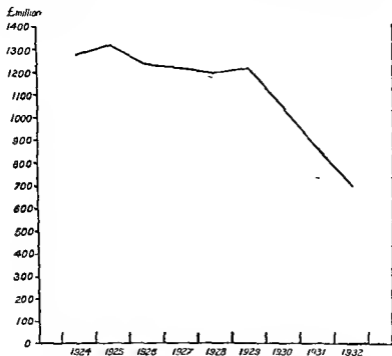


DIAGRAM 15 Value of Imports into United Kingdom

for the loss suffered by replacing precise figures by a graph which cannot pretend to delicate precision at all

When we are plotting the points on the graph corresponding to the values of the variable quantity we bear in mind two things: (1) the fact that we wish to have a graphical representation corresponding exactly to the

original data, (2) the possibility that what we are really interested in is the extent of the changes which are taking place in our variable. In (12) and (13) we have graphical representations corresponding exactly to the original data. Sometimes we may be more interested in the changes which take place, rather than in the actual size of the figures involved. If this is the case, we may need a larger

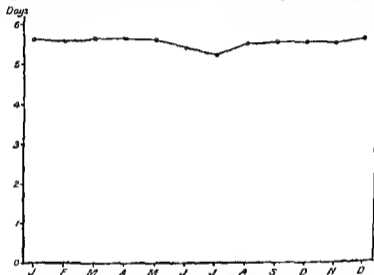


DIAGRAM 16. Number of Coal-winding Days per week

vertical scale than is consistent with the size of the paper on which the diagram is being made, and we therefore sacrifice a part of the diagram by elimination. For instance, suppose we wished to show graphically the following figures —

AVERAGE WEEKLY NUMBER OF DAYS IN WHICH COAL WAS WOUND IN  
ONE FORTNIGHT OF EACH MONTH OF 1913

Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
5.64	5.61	5.67	5.69	5.64	5.44	5.26	5.54	5.60	5.59	5.56	5.66

These are shown in Diagram 16, plotted in the usual way, and in Diagram 17 on a larger scale to emphasize the changes which take place from month to month, but as this scale would involve the zero of the vertical scale, with the horizontal scale, placed some considerable distance from the graph, we eliminate that part, and indicate this gap as in the diagram. Quite often, however, this indica-

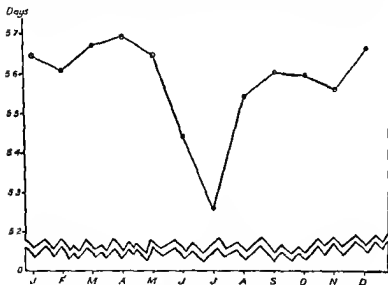


DIAGRAM 17 Number of Coal winding Days per week

tion of the gap is not shown in the diagram at all, the only means of noting it is by observing that the zero on the vertical scale does not coincide with the horizontal base line on which the time element is shown

This method of reducing the space occupied by the diagram, when a large scale is used for the variable which is being plotted, is very usual, especially where expense

of printing enters into consideration, but when it is adopted care must be taken in reading the diagram. There is no doubt that the most important part of the diagram to impress itself on a person reading it is the graph itself. He may neglect to observe the vertical scale, or he may only take note of that after he has had the changes shown by the graph firmly impressed on his mind. This first impression may linger even though it may be somewhat modified by a sight of the scale. It is quite conceivable

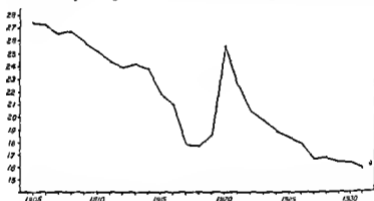


DIAGRAM 18 Birth-rate England and Wales (Births per thousand of population)

that the total impression received from such a diagram may be substantially different from that obtained if the diagram had been differently constructed. Thus Diagram 18, showing changes in the birth-rate since 1905, indicates a considerable decline in the last twenty or thirty years, interrupted by the War years and those immediately succeeding. This decline, it is true, has been considerable, but is not as great as is suggested on a first reading of this diagram. If the figures had been plotted on a different

scale, where the zero on the birth-rate scale is shown, the first impression would be quite different.

Diagrams are not only used to show single time series, but to enable comparisons to be made with other series. There are two kinds of such diagrams, the first are those where the same units are involved in the two series, and where the same scale will serve for both graphs or indeed

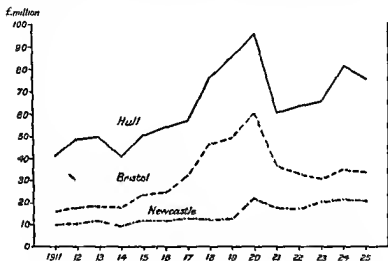
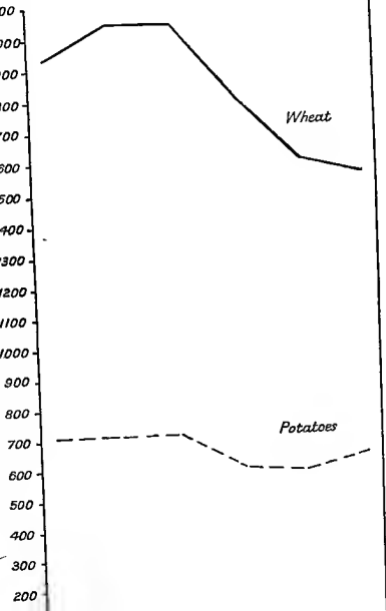


DIAGRAM 19 Value of Imports of Merchandise at Hull, Bristol, and Newcastle

for three or four graphs, if more than two series are being represented graphically in the same diagram; the second are those where different units are involved in the two or more series. In the first kind no difficulties apparently arise, though in particular cases such a diagram may not properly succeed in presenting the comparable facts in a proper manner. It must be remembered that even with

and  
es



a simple series there are two aspects to be considered in graphical representation, (1) the graph is to show properly the actual sizes of the numbers in the series, (2) the graph is to show properly changes in these, so that comparisons can be made. If we have two series in a diagram, there are six aspects to be considered, for each series the sizes of the numbers and the changes in them, and between the series a comparison of the sizes of the numbers and a comparison of the changes. So if we have three series to be plotted,

OUTPUT OF CERTAIN MINERALS IN GREAT BRITAIN  
(IN MILLION TONS)

	Coal	Iron Ore and Iron Stone	Tin Ore, dressed
1920	230	13	—
1921	163	3	—
1922	250	7	—
1923	276	11	—
1924	267	11	—
1925	243	10	—
1926	126	4	—
1927	251	11	—
1928	237	11	—

there are twelve points of view to be borne in mind. It is quite conceivable that a diagram may satisfactorily succeed in showing the series from some aspects but may fail in other respects. Diagram 19 showing value of imports into three ports reasonably serves its purpose from all points of view, but Diagram 20 is on such a scale that changes in acreage of beet crops are hardly discernible. On the other hand this diagram adequately represents the various acreages of these different crops. It would be necessary, if we wished to give a better graphical

representation of the acreage under beet, to show this separately in another diagram. This kind of breakdown is of course no different from a similar breakdown in certain tables. For instance, we might represent the facts relating to output of certain minerals in Great Britain in the table on p. 185.

The output of tin ore is small, and does not run into millions of tons, the actual details are :—

OUTPUT OF CERTAIN MINERALS IN GREAT BRITAIN (TONS)

	Coal	Iron Ore and Iron Stone	Tin Ore, dressed
1920	229,532,081	12,677,670	4,858
1921	163,251,181	3,470,518	1,078
1922	249,606,864	6,836,507	650
1923	276,000,560	10,875,211	1,760
1924	267,118,167	11,050,589	3,547
1925	243,176,231	10,142,878	4,032
1926	126,278,521	4,094,388	3,578
1927	251,232,336	11,206,601	4,321
1928	237,471,931	11,262,323	4,844

Actually, of course, if we wished to show these figures in round numbers, we should make a table :—

OUTPUT OF CERTAIN MINERALS IN GREAT BRITAIN

	Coal (mn. tons)	Iron Ore and Iron Stone (mn tons)	Tin Ore (thousand tons)
1920	230	12·7	4·9
1921	163	3·5	1·1
1922	250	6·8	0·7
1923	276	10·9	1·8
1924	267	11·1	3·5
1925	243	10·1	4·0
1926	126	4·1	3·9
1927	251	11·2	4·3
1928	237	11·3	4·8

This change in the degree of precision of the round numbers in the table corresponds, of course, to change of scale in a diagram.

When many time series are represented graphically on the same diagram, some confusion may arise because the graphs may be nearly superimposed on one another, or because they may cross and recross. Even though we carefully differentiate between them by means of broken and dotted lines or with different coloured inks, it may

QUANTITIES OF WHEAT (GRAIN) CONSIGNED FROM CERTAIN COUNTRIES  
TO THE UNITED KINGDOM  
(Million Cwts)

	Russia	Argentina	U S A.	Canada	British India	Australia
1905	25.8	23.3	6.5	6.8	22.8	10.1
1906	16.1	19.2	22.6	11.2	12.8	7.9
1907	11.4	21.9	19.9	13.2	18.3	8.3
1908	5.1	31.7	25.8	15.8	2.9	5.5
1909	17.8	20.0	15.5	18.6	14.8	9.7
1910	28.9	15.1	10.9	18.4	17.9	13.1
1911	18.1	14.7	12.9	14.4	20.2	13.9
1912	9.0	18.8	20.0	21.6	25.4	11.9
1913	5.0	14.8	34.1	21.8	18.9	10.1

be difficult for a person to appreciate the facts which it is the purpose of the diagram to convey. There is a point at which the diagram of this kind becomes so confusing that it serves no useful purpose at all. Any one proposing to understand it may quite likely give up the attempt after a first glance at the diagram. Obviously such a diagram is of no use. A guiding principle to remember in the construction of diagrams is that they are to serve to help others to appreciate the facts contained in certain tables, the idea being that most persons can more readily

appreciate these when they are presented in diagrammatic form.

Diagram 21 shows the figures of the table on p. 187.

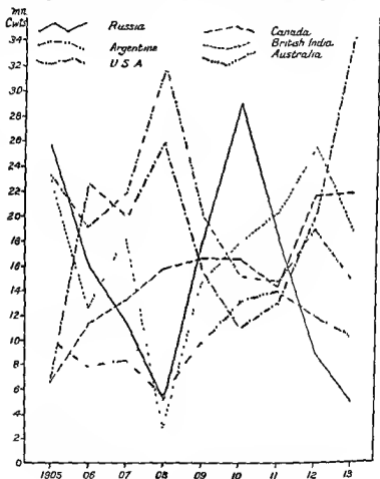


DIAGRAM 21 Quantities of Wheat (Grain) consigned from Certain Countries to United Kingdom

It is questionable whether such a diagram is a real help to anyone interested in appreciating this table.

We have now to consider the diagrammatic representation of two or more series of figures which involve different units, so that more than one vertical scale is necessary in the diagram. Difficulties arise here because the impression conveyed by the diagram depends upon the scales used, and these may be determined merely by considerations suggested on account of the necessity of obtaining a diagram free from obvious defects such as confusion of the graphs in it. Considerations of this kind might result in a number of methods of plotting the figures, which would lead to diagrams of different appearance.

Suppose we wished to make a diagram to show these figures :—

GREAT BRITAIN : COAL MINING INDUSTRY OUTPUT and MAN-SHIFTS  
WORKED QUARTERLY FIGURES 1922 to 1924

	Quarters	Output (mn tons)	Man-shifts worked (mns.)	Output per Man-shift (cwt.)
1922	1	57.6	63.2	18.23
	2	53.3	59.8	17.80
	3	58.7	65.4	17.94
	4	64.5	71.3	18.10
1923	1	67.1	73.5	18.25
	2	65.5	73.2	17.90
	3	62.0	71.2	17.42
	4	67.8	76.4	17.76
1924	1	67.0	75.4	17.79
	2	61.6	70.4	17.48
	3	59.2	68.3	17.33
	4	62.4	70.4	17.74

Diagrams 22 and 23 are only two different methods of presenting these three series. From Diagram 22 we

should get the impression that output per man-shift hardly changed at all during this period, and that the changes in the other two series were nearly of the same extent.

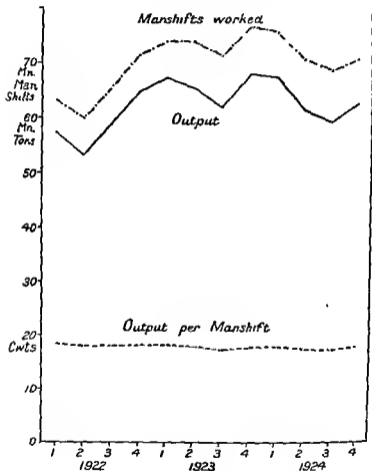


DIAGRAM 22 Great Britain. Coal Mining

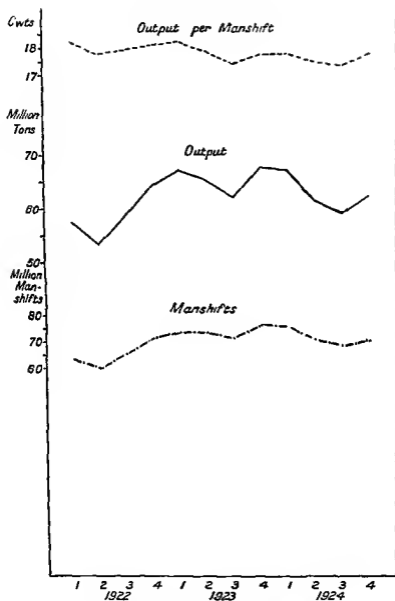


DIAGRAM 23. Great Britain. Coal Mining.

From Diagram 23 we should be impressed by the fact that changes of a regular character exist in all these series, and that those in output per man-shift were of the same degree as those in the number of man-shifts worked. So long as the guiding principle of making a diagram free from confusion is observed there are no rules which can be laid down for the construction of such diagrams in which different units are involved in the series.

COAL MINING INDUSTRY · OUTPUT AND MAN-SHIFTS WORKED  
(Relative figures)

		(a)			(b)		
		Output	Man-shifts worked	Output per Man-shift	Output	Man-shifts worked	Output per Man-shift
1922	1	100	100	100	84.9	82.7	102.6
	2	92.5	94.6	97.6	78.6	78.2	100.2
	3	101.9	103.5	98.4	86.6	85.6	101.0
	4	112.0	112.8	99.3	95.1	93.3	100.9
1923	1	116.5	116.3	100.1	99.0	98.2	102.8
	2	113.7	115.8	99.2	96.6	95.6	100.6
	3	107.6	112.7	95.6	91.4	83.2	98.1
	4	117.7	120.9	97.4	100	100	100
1924	1	116.3	119.3	97.6	98.8	98.7	100.2
	2	106.9	111.4	95.9	90.9	92.1	98.4
	3	102.8	108.1	95.1	87.3	89.4	97.6
	4	108.3	111.4	97.3	92.0	92.1	99.9

One of the simplest methods of avoiding this difficulty of choice of scales is to sacrifice the units involved and resort to percentages. This means, of course, that the actual sizes of the figures in the series are not reproduced graphically at all, that only the relative figures are shown in the graphs, and we therefore do not pretend that the diagram is making any attempt to show the original table properly. But even here we have an embarrassing

choice to make. Shall we take each figure in a series as a percentage of the first figure in the series, or the last figure, or the maximum figure, or the minimum, or the

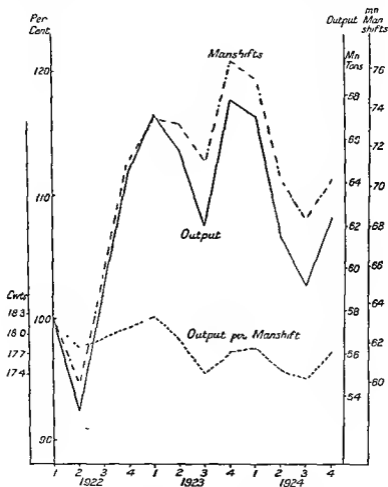


DIAGRAM 24. Great Britain. Coal Mining.

average of the series? Each of these different methods would be useful and appropriate on different occasions, but whichever we use, we have now the advantage that we have eliminated altogether the question of scales for

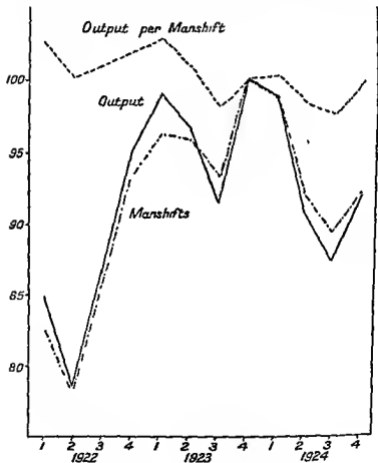


DIAGRAM 25. Great Britain. Coal Mining.

the graphs, because now all our series no longer involve units. Let us see the result of taking each series in the last table as (a) percentages of the first figure in the series, (b) as percentages of the figure for 1923, fourth quarter, which is the maximum figure in each of the first two series. We shall have to plot the figures in the table on p. 192.

These are shown in Diagrams 24 and 25.

The zero on the percentage scale is not shown, as it has no particular significance in our diagrams, indeed we may consider that 100 per cent is the real basic figure in these series. This may be emphasized by drawing a horizontal line through 100 per cent on the vertical scale. We could, of course, superimpose on these diagrams the scales corresponding to the original units. We have taken, in Diagram 24, 57.6 million tons of output, 63.2 million man-shifts, and 18.23 cwt. per man-shift as 100 units in each case, thus we have these percentages corresponding to certain figures of output, man-shifts and, output per man-shift :—

Output		Man-shifts		Output per Man-shift	
Mn tons	Per-centage	Mns	Per-centage	Cwts	Per-centage
52	90.3	60	94.9	17.3	94.9
54	93.7	62	98.1	17.4	95.4
56	97.2	64	101.3	17.5	96.0
58	100.7	66	104.4	17.6	96.5
60	104.2	68	107.6	17.7	97.1
62	107.6	70	110.8	17.8	97.6
64	111.1	72	114.0	17.9	98.2
66	114.6	74	117.1	18.0	98.7
68	118.1	76	120.3	18.1	99.3
				18.2	99.8
				18.3	100.4

These equivalents in the percentage scale enable us to show the appropriate scales of values on the percentage diagram. This has been done in Diagram 24. Naturally the zeros in these scales all coincide, but they are not shown in the diagram.

### *Ratio or Logarithmic Scales*

Another method of dealing with the graphical representation of series involving different units is to plot the figures on a ratio scale and avoid the trouble occasioned by the calculation of the percentages. In effect this means departing altogether from the ordinary kind of scale and the introduction of a new idea in graphical methods. In the usual diagram which has so far been considered the same length on the paper in any part of the scale is equivalent to the same number of units. Thus if the scale is, 100 tons equals 1 inch, then the interval on the scale between 500 and 600 tons is the same (1 inch) as the interval on the scale between 200 and 300 tons. In a ratio scale this is not the case, the length of the interval between two values on the scale is proportional to the ratio between these two values.

RATIO SCALE



On this scale the distance between 1 and 2 is the same as that between .5 and 1 or 2 and 4 or 4 and 8. In each case the ratio between these pairs is 2 : 1. Similarly the same distance is observed between 1 and 3, 3 and 9, 4 and 12. Consecutive points on the scale corresponding to consecutive integers get closer and closer together as we

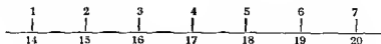
increase the numbers from 1. The simplest method of obtaining such a scale is to determine the position of the number on the scale from its logarithm. The table of logarithms below was used in the construction of the scale above.—

Number	.5	1	2	3	4	5	6	7	8	9	10	11	12
Log	— 301	0	301	477	602	699	778	845	903	954	1 0	1 041	1 079

A logarithm of 1 was taken as 1 cm. in the scale. Thus .301 is 3.01 cm., .477 is 4.77 cm., and so on. The actual distance between the point marked 1 on the scale and that marked 12 is 10.79 cm. The distance between the point marked 6 on the scale and that marked 12 is the difference between 7.78 cm. and 10.79 cm., corresponding to the difference between log 6 and log 12, this difference is, of course, log 2, and the distance between the points marked 4 and 8 on the scale, the difference between 6.02 cm. and 9.03 cm. is the same amount. We are using here the well-known formula

$$\text{Log } X - \text{Log } Y = \text{Log } (X/Y)$$

In any ordinary scale the same marks on the scale can be taken to correspond to different groups of numbers by adding the same amount to each figure on the original scale, as below :—



Here the same marks on the scale will correspond to the group of figures 1, 2, 3, 4, . . . or to the group 14, 15, 16, 17, . . . or to any other group obtained by adding the same amount to each of the original figures. So, in a ratio

scale, the same marks on the scale can be taken to correspond to different groups of numbers by multiplying each figure in the original scale by the same amount:—

1	2	3	4	5	6	7	8	9	10
3	6	9	12	15	18	21	24	27	30
.4	.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0

Thus the points on the scale corresponding to 1 to 10 may also be taken as 3 to 30, or .4 to 4.0, and so on.

Ratio scale graph paper can therefore easily be made or printed, just as ordinary squared paper is made, and graphs can be plotted on this kind of paper just as on the usual graph paper, except that now the graphs are on a ratio or logarithmic scale instead of on an ordinary scale. This kind of graph paper is often referred to as "logarithmic paper". There is no zero on logarithmic paper, as  $\log 0$  is an indefinitely large negative number.

On such paper as this, graphs can be traced in whatever units are involved without any difficulty, and the resulting diagrams will show both the actual figures given in the tables and at the same time will indicate relative values, both between numbers in the same series, and corresponding numbers in different series. A change between consecutive figures in a series indicated on such a diagram by a certain length will correspond to a relative change of the corresponding amount. The change from A to B in Diagram 26 is the same as that between C and D, in each case an increase of 33.3 per cent is recorded.

If printed logarithmic paper is not available for use on any occasion, a logarithmic scale diagram can be constructed by plotting the logarithms of the numbers in the series instead of the actual numbers, on ordinary

graph paper, and the original units can be shown in the diagram at the same time. For instance let us take the table of figures relating to output and man-shifts worked

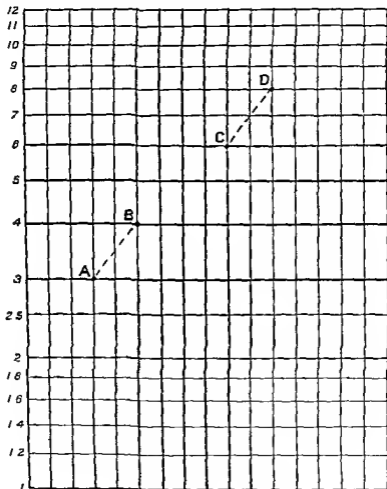


DIAGRAM 26 Logarithmic Paper

in the coal mining industry previously quoted on p. 189; we look up the logarithms of the numbers in the table and plot the latter.

LOGARITHMS

	Quarter	Output	Man-shifts	Output per Man-shift
1922	1	1.760	1.801	1.261
	2	1.727	1.777	1.250
	3	1.769	1.816	1.254
	4	1.810	1.853	1.258
1923	1	1.827	1.866	1.261
	2	1.816	1.865	1.253
	3	1.792	1.852	1.241
	4	1.831	1.883	1.249
1924	1	1.826	1.877	1.250
	2	1.790	1.848	1.243
	3	1.772	1.834	1.239
	4	1.795	1.848	1.249

We construct a scale for logarithms ranging from 1.2 to 1.9 and plot the series in *Diagram 27*. We can arrange scales of output, etc., corresponding to the appropriate figures on the logarithmic scale. As we are not interested in the logarithms of these numbers except in so far as they enable us to construct a diagram, the logarithmic scale in the diagram is eliminated, leaving only the scale showing the original unit involved. In this way we construct the diagram on a logarithmic or ratio scale without having the appropriately ruled paper.

Obviously the position of one graph relative to another in such a diagram is a mere accident dependent on the units in the original table. Thus output might have been quoted in the table in cwts., in which case the characteristics of the logarithms of output would have

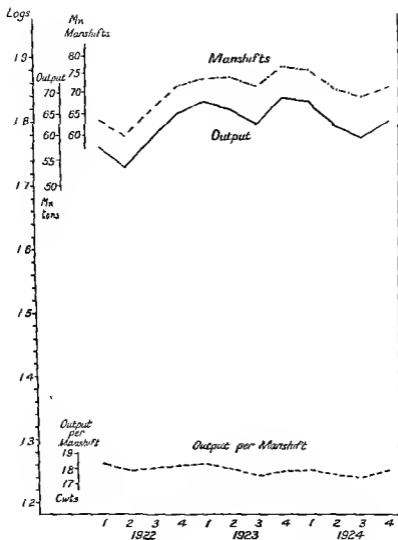


DIAGRAM 27. Coal Mining Figures on Logarithmic Scale

been 3, or output per man-shift might have been given in tons instead of cwts., in which case all the logarithms of this series would have been negative. The resulting appearance of the graphs would be unchanged, because the differences

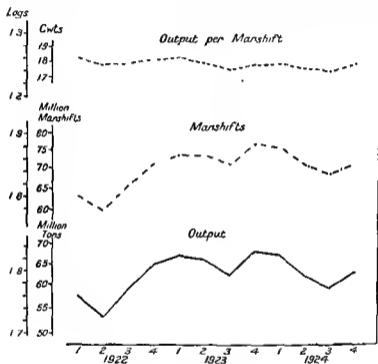


DIAGRAM 28

between the logarithms of the series would remain the same, although the original series might have been multiplied or divided by a constant amount. This is equivalent to saying that points on the scale of logarithms shown as 1.2, 1.3, 1.4, 1.5, etc., could equally be shown

as 3·2, 3·3, 3·4, 3·5, etc., or 1·7, 1·8, 1·9, 2·0, etc., or any other series obtained from the first by the addition to each of a constant amount. Diagram 28 shows these same three series as before on a logarithmic scale. The graphs have changed relative positions in the diagram, but are otherwise unaltered.

Naturally, if we changed the scale of the logarithms from one graph to another the appearance of the diagram

UNITED KINGDOM RECEIPTS FROM SUPER TAX

Year ending 31st March	£000	(logs)
1911	2,891	3.46
1912	3,018	3.48
1913	3,600	3.56
1914	3,339	3.52
1915	10,121	4.01
1916	16,788	4.23
1917	19,140	4.28
1918	23,279	4.37
1919	35,560	4.55
1920	42,405	4.63
1921	55,669	4.75
1922	61,351	4.79
1923	63,910	4.81
1924	61,747	4.79
1925	62,969	4.80
1926	67,833	4.83

would not be unaltered. If 1 inch is taken to correspond to 0·1 in a scale of logarithms, then the distance of the point on this scale corresponding to 1·9 from that corresponding to 1·8 must be 1 inch; it is immaterial where the point corresponding to 1·8 is placed, on the line showing the scale of logarithms. If the scale is altered to 2 inches to 0·1 in a scale of logarithms, then of course the distance between two such points would now

be 2 inches and the appearance of any graph would be altered.

A logarithmic scale is appropriately used also when we

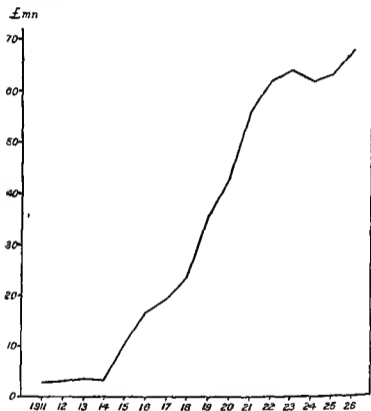


DIAGRAM 29 Super Tax Receipts

wish to graph a series of figures which change fundamentally as time goes on, increasing or decreasing at a great rate. If an ordinary scale were used the smaller figures would

hardly be distinguishable from each other in the graph, because the scale would be small in order to include the larger figures. But a logarithmic scale has the apparent effect of magnifying the scale in its lower part and contracting it in its upper regions, and on such a scale a series of this kind would be satisfactorily plotted. As

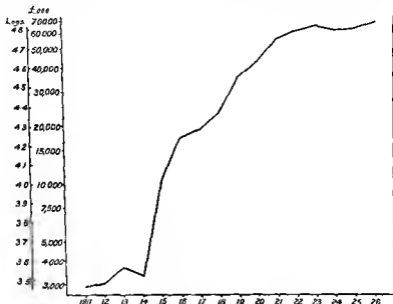


DIAGRAM 30 Super Tax Receipts (logarithmic scale)

an illustration let us consider the figures relating to revenue from super tax. (Table on p. 203)

These figures are represented graphically on an ordinary scale and on a logarithmic scale in Diagrams 29 and 30 respectively. The changes between the years 1911-14 are hardly noticeable in 29 but are magnified in 30, the considerable relative change between 1914 and 1915 is

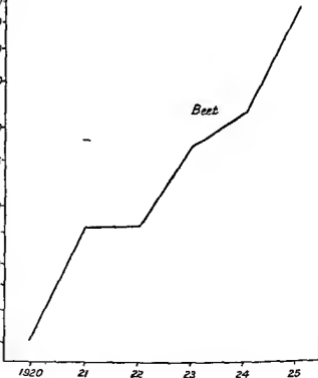
Logs Million  
Acres4  
3  
2  
12.25  
2.0  
1.75  
1.5*Wheat*19  
8  
7  
6  
5  
4  
3  
2  
1  
0  
Thousand  
Acres60  
50  
40  
30  
20  
15  
10  
8  
6  
5  
4  
3*Beet*

DIAGRAM 31 Great Britain. Acreage of Crops.

given due prominence in 30 but not in 29 and the decline in the rate of increase in later years is brought out in 30 whereas 29 emphasizes the actual increments each year.

As a further illustration of a slightly different kind we may instance the graphing of the following figures relating to acreage of wheat and beet 1920-5, which were shown on an ordinary scale in Diagram 20.

GREAT BRITAIN      ACREAGE OF CROPS      1920-5

	Wheat (000 acres)	Beet (Acres)
1920	1,929	3 045
1921	2,041	8,334
1922	2,032	8,413
1923	1,799	16 923
1924	1,594	22,637
1925	1,548	56,243

These figures are shown on a logarithmic scale in Diagram 31, and the graphs present an entirely different appearance from the corresponding graphs in Diagram 20.

## CHAPTER 12

### ANALYSIS OF TIME SERIES

When we consider the changes with time of a certain quantity we are concerned to interpret these changes, and to observe how they are related to similar changes which are apparent in other time series. For instance when we examine the series of figures below giving output of coal in Great Britain, we naturally ask ourselves to what the changes taking place are due, and how they are related to changes in other series.

OUTPUT OF COAL GREAT BRITAIN, 1873-1928

	Output (million tons)		Output (million tons)		Output (million tons)		Output (million tons)
1873	129	1887	162	1901	219	1915	253
1874	127	1888	170	1902	227	1916	256
1875	133	1889	177	1903	230	1917	248
1876	134	1890	182	1904	232	1918	228
1877	134	1891	185	1905	236	1919	230
1878	133	1892	182	1906	251	1920	230
1879	134	1893	164	1907	268	1921	163
1880	147	1894	188	1908	262	1922	250
1881	154	1895	190	1909	264	1923	276
1882	156	1896	195	1910	264	1924	267
1883	164	1897	202	1911	272	1925	243
1884	161	1898	202	1912	260	1926	126
1885	159	1899	220	1913	287	1927	251
1886	158	1900	225	1914	266	1928	237

We observe that there is, on the whole, a gradual increase in output of coal, that there are sudden breaks in this

increase of large and small amounts. We are aware that these figures are the results of the combined operations of management and workers in the Coal Mining Industry. We realize that they are stimulated to greater output by increasing demand for their product. We know that this demand is not constant, we know also that industrial troubles from time to time have the effect of stopping output temporarily, accidents put a mine out of operation, a mine ceases to be a valuable proposition at the prices then current, new mines are opened, technical advances enable coal to be won more easily, and so on. We may say that the figures in this series result from the operation of a large number of causes of different kinds, and we must consider the nature of these causes, in order that we may determine, if possible, the effect of them, and so that we may relate the size of the effect produced by a group of causes on one phenomenon to the similar effect on another phenomenon.

Let us first consider the kind of causes which are operating to produce certain effects. First, we must reach down to fundamental facts. At a given time there is a certain number of human beings in the world, a certain amount of habitable and cultivated land, a certain number of domesticated animals in the service of the human beings. As time progresses there is a gradual change in the land under cultivation, a gradual change in the number of animals. These changes in recent historical times have all been in the nature of increases, and there have been consequently gradual increases in demand for certain commodities with resultant increases in the supply of them. We may therefore consider that there is a certain *growth* factor which is operating to produce gradual

changes in a certain series. The resultant gradually changing nature of the series is generally referred to as the *secular trend* of the series, and this trend must be considered as linked up to the growth factor referred to above.

Secondly, operating at the same time as the growth factor, there is a *group of causes* which do not operate continuously, but in a regular spasmodic manner. Thus, the seasons recur in the same way each year, day follows night with regularity. In certain areas of the earth there is a wet season followed by a dry season, certain ports are regularly frozen up in the winter and as regularly free in the summer. Seed is sown in spring and the crop is reaped in autumn. The result of the operation of causes of this kind is a regular up and down movement in a series of figures relating to some phenomenon, which has been observed, and which is *affected* by this group of causes. This movement is generally referred to as the *seasonal movement*. If the output of coal quarter by quarter in Great Britain were observed, an up and down movement of this kind would be noted, due to the changing demand for coal in winter and summer, and this movement would be superimposed on the general trend already referred to.

During the nineteenth century a similar fairly regular up and down movement has been observed in a large number of time series of economic data, these movements being repeated at intervals of 7-11 years. The causes of these periodic or cyclical changes are in doubt, but there is no doubt about their existence. Those years when the observed phenomena show upward movements are referred to as "boom" years and those years of downward movements as years of "depression" or

"crisis". The typical up and down movement of this kind is referred to as the "trade cycle". The determination of the extent and regularity of these movements is one of the matters we must concern ourselves with in our analysis of time series.

In addition to the group of causes operating to produce regular up and down movements in our series, there is another group which operates in an adventitious manner. This group of causes includes such events as floods completely ruining a particular area, its crops and houses, and resulting in the deaths of many of its people; strikes, resulting in a cessation for long or short periods of production; deaths of monarchs which might put a stop to certain events which normally would occur; fires and earthquakes, wars and revolutions, and so on. All these causes operate from time to time, there is no regularity in their operation, the effects may be large or small, but they certainly exist. In this group also would be considered the adventitious element in the gradual growth factors or in the regularly operating spasmodic factors. Thus by a peculiar combination of wind, sunshine, and rain in a certain season there may be a bumper crop or a poor crop of some agricultural product. Or, an invention or discovery may hasten the gradual growth effect, and change the nature of a certain series fundamentally.

In any given time series, then, we look for three kinds of movement :—

- (1) General trend.
- (2) Regular fluctuations
  - (a) Seasonal.
  - (b) Cyclical.
- (3) Irregular fluctuations.

We attempt to analyse a series into these three constituent parts. When considering the relationship between one series and another we attempt to relate each corresponding part of the two series.

A given series may be composed of all three kinds of movement, or it may not fluctuate at all. There may be a general trend upwards or downwards, or the general size of the figures in the series may be the same as time goes on. We will consider the general case where all three elements are supposed to enter.

Our analytical problem can be illustrated by considering how a given series may be made up. Let us suppose we are dealing with an annual series in which the seasonal movement (if any) will be concealed, but which is made up of a general trend, cyclical fluctuations and irregular fluctuations. Let us suppose that the series is constructed as in the table on p. 213.

Here we have a series in column 5 which has been made up as shown in the table. The problem we have to consider in practice is, being given only such a series as that in column 5, can we reconstruct columns 2, 3, and 4? We realize that, whatever solution is obtained as a result of the analysis, this solution will necessarily be approximate only, but such an approximate solution may be sufficient for practical purposes in particular cases.

The simplest method of analysis is, first to eliminate entirely, as far as is possible, all the fluctuations from the series, whether of a regular nature or an irregular nature, leaving us only with the general trend. Having this, we can now obtain the total fluctuation in the series for each year, since for any given year the value of the series is equivalent to the trend value plus the fluctuation. The

fluctuations can then be analysed separately in an endeavour to find the extent of the regular part of the fluctuations. When this is known simple subtraction gives us the irregular part of the fluctuations. The procedure outlined above will be considered in more detail now.

(1) Year	(2) General Trend	(3) Cyclical Fluctuations	(4) Irregular Fluctuations	(5) Resulting Series
1	10.0	+ 1.5	- 0.4	11.1
2	10.1	+ 1.0	+ 2.0	13.1
3	10.2	0	- 1.9	8.3
4	10.3	- 1.0	+ 0.7	10.0
5	10.4	- 1.5	+ 1.2	10.1
6	10.5	- 1.0	- 0.3	9.2
7	10.6	0	+ 0.8	11.4
8	10.7	+ 1.0	- 0.2	11.5
9	10.8	+ 1.5	- 0.6	11.7
10	10.9	+ 1.0	+ 0.4	12.3
11	11.0	0	0	11.0
12	11.1	- 1.0	- 0.7	9.4
13	11.2	- 1.5	+ 0.3	10.0
14	11.3	- 1.0	0	10.3
15	11.4	0	+ 1.6	13.0
16	11.5	+ 1.0	- 1.1	11.4
17	11.6	+ 1.5	- 0.8	12.3
18	11.7	+ 1.0	+ 1.5	14.2
19	11.8	0	+ 0.8	12.6
20	11.9	- 1.0	- 0.8	10.1
21	12.0	- 1.5	+ 1.9	12.4
22	12.1	- 1.0	- 0.4	10.7
23	12.2	0	+ 0.7	12.9

The elimination of the fluctuations can be done by making a graphical representation of the given series, observing the serrated nature of the graph, and smoothing out the irregularities by drawing a freehand curve which appears satisfactorily to describe the general trend of the graph. This method has the advantage of quickness of

execution, but is unsatisfactory in that it leaves the determination of the position of the trend to the individual, thus two persons might arrive at different conclusions as to the nature of the trend values, because they have drawn two different smoothed curves. On the other hand, the fact that each of them has made his best endeavour to arrive at the best result, and these differing, emphasizes the non-precise nature of the conclusion. As we pointed out, any result which is obtained is necessarily approximate. The table above has been



DIAGRAM 32

graphed in Diagram 32 and the trend has been drawn in what appears at first sight to be the correct position. The value on the trend in the first year is 9.6 and in the last year 12.6, these figures should be contrasted with 10.0 and 12.2 of the table. It is likely that others trying to estimate the trend from the graph would arrive at slightly different conclusions from those shown in Diagram 32

A more usual method of elimination of the fluctuations is by using moving averages. We will proceed to disclose this method, and what it involves.

We are given a series of values of some quantity at

equal time intervals. Groups of  $n$  successive values of the series are averaged, these groups being composed as follows, the first group consists of the first  $n$  items of the series, the second group consists of the items from the 2nd to the  $(n + 1)$ th, the third group consists of the items from the third to the  $(n + 2)$ th, and so on. These averages are supposed to give the trend values corresponding on the time scale to the middle period between the first and  $n$ th time intervals, the middle period between the second and  $(n + 1)$ th time intervals, and so on. The choice of the  $n$  values to be averaged is in the hands of the operator and is arbitrary, though in practice he would be guided by certain considerations to be discussed later. Suppose for example we considered the made-up series in the table above, and calculated 5-yearly moving averages. The average of the first five items in the series is 10.3 and represents the trend value for the mid-year of the first five, i.e. year 3.

Year	Series	Sums	Averages
1	11.1		
2	13.1		
3	8.3	51.7	10.3
4	10.0	52.0	10.4
5	9.2	50.4	10.1
6	11.4	53.8	10.8
7	11.5		
8	11.7		

So, the average of the five items from the second to the sixth is 10.4 and gives the trend value for the year 4, half-way between the second and sixth years, and so on.

These averages, which are supposed to give the trend values, are called Moving Averages.

We must naturally consider how far this method does in fact perform the function expected of it. It is supposed to operate on a given series and eliminate the fluctuations, and give only the general trend. Therefore, if this process is used on a series which does not contain fluctuations at all, i.e. a trend series only, it should reproduce the original series exactly. Also, if this process is used on a series consisting only of fluctuations these should be entirely eliminated, leaving us with a series of zeros. Let us first of all consider a series of figures without fluctuations which, when plotted, would be graphically represented by means of a straight line.

Time Interval	Series	Sums in 5's	5-Interval Moving Average	Sums in 7's	7-Interval Moving Average	Sums in 8's	8-Interval Moving Average
1	1						
2	3						
3	5	25	5				
4	7	35	7	49	7	64	8
5	9	45	9	63	9	80	10
6	11	55	11	77	11	96	12
7	13	65	13	91	13	112	14
8	15	75	15	105	15	128	16
9	17	85	17	119	17	144	18
10	19	95	19	133	19	160	20
11	21	105	21	147	21	176	22
12	23	115	23	161	23		
13	25	125	25				
14	27						
15	29						

If these are averaged, however many are grouped, the resulting averages are exactly the same as the original figures, when an odd number of items are taken; or if an even number are averaged the results are trend values corresponding to positions on the time scale half-way between the given time intervals. Thus the 8-interval

moving averages give trend values for the time intervals 4·5, 5·5, 6·5, etc.

This simple illustration is merely an example of a general principle:—If the moving average process operates on a series of values, which when plotted are shown by a straight line, then the result of the process is to reproduce the original series, or another series which, when plotted, give points on the original line. This is true whether the line shows increasing, decreasing, or stationary values in the series. A straight line trend is reproduced exactly by this process.

On the other hand, a curved trend is not exactly reproduced by the process. Let us illustrate with a series obtained by giving  $x$  successive values 1, 2, 3, . . . in the expression  $x + \frac{1}{2}x + \frac{1}{2}x^2$ . The table below shows the working of the 5-, 6-, 7-interval moving averages process.

Time Interval	Series	Sums in 5's	5-Interval Moving Average	Sums in 6's	6-Interval Moving Average	Sums in 7's	7-Interval Moving Average
1	2						
2	4						
3	7	40	8	62	10·3	91	13
4	11	60	12	89	14·8	126	18
5	16	85	17	122	20·3	168	24
6	22	115	23	161	26·8	217	31
7	29	150	30	206	34·3		
8	37	190	38				
9	46						
10	56						

In the case of the 5- and 7-interval moving averages these definitely are not the same as the original series. The 6-interval moving averages also are not the same as the values of the original expression obtained when  $x$  is given values  $3\frac{1}{2}$ ,  $4\frac{1}{2}$ ,  $5\frac{1}{2}$ ,  $6\frac{1}{2}$ ,  $7\frac{1}{2}$ , these being 8·9, 13·4,

18.9, 25.4, 32.9. The process gives figures which are greater than those in the original series.

Let us now consider a series obtained by giving  $x$  successive values 1, 2, 3, . . . in the expression  $16 + 3\frac{1}{2}x - \frac{1}{2}x^2$ , and operate on the series with 5- and 7-interval moving averages.

Time Interval	Series	Sums in 5's	5-Interval Moving Average	Sums in 7's	7-Interval Moving Average
1	19				
2	21				
3	22	105	21		
4	22	105	21	140	20
5	21	100	20	133	19
6	19	90	18	119	17
7	16	75	15	98	14
8	12	55	11		
9	7				
10	1				

Again, the process does not reproduce the original series, here all the results are less than the given figures. These two examples are illustrations of the working of a general principle —when the moving average process operates on a trend which is graphically represented by a curve, the result is another curve different from the original. If the trend curve is convex to the horizontal time axis, the moving average curve is above the original; if the trend curve is concave to the time axis the moving average curve is below the original. More generally, the process tends to reduce the curvature in the original graph. The moving average curve will be closer to the original curve in that part of it where the curvature is least, and where a curve is very nearly equivalent to a straight line, the moving average curve will be practically coincident with it.

Let us take another illustration where the original curve has varying curvature in different parts. Diagram 33 shows

Time Interval	Series	5-Interval Moving Average
1	37	
2	59	
3	75	70
4	86	82
5	93	90
6	97	95
7	99	98
8	100	100
9	101	102
10	103	105
11	107	110
12	114	118
13	125	130
14	141	
15	163	

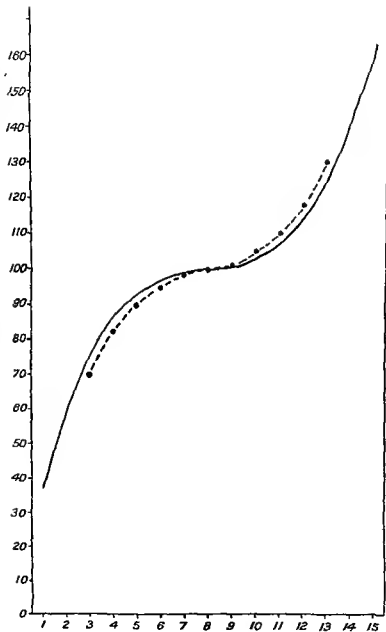
the original series together with the result of the 5-interval moving average process operating on the series. Where the original series is concave to the time axis, the moving average curve is below, and where the original is convex, the moving average is above it. Also, the process gives results which are closest to the original where the curvature is not so marked. The difference between the two is five when the time interval is 3 and 13, and is one when the time interval is 7 and 9.

Thus the process introduces distortion when the original series is curved. This is the great disadvantage of this method. On the other hand the amount of distortion is not great in those cases where there is not a large amount of curvature in the original.

Let us now consider the effect of using the method on a series of regular fluctuations.

Time Interval	Series	Sums in 5's	5-Interval Moving Average	Sums in 7's	7-Interval Moving Average	Sums in 9's	9-Interval Moving Average
1	+ 2						
2	+ 2						
3	0	+ 1	+ .2				
4	- 1	- 3	+ .6	0	0		
5	- 2	- 4	- .8	0	0	+ .4	+ .4
6	- 2	- 2	- .4	0	0	+ .2	+ .2
7	+ 1	+ 1	+ .2	0	0	- .1	- .1
8	+ 2	+ 3	+ .6	0	0	- .3	- .3
9	+ 2	+ 4	+ .8	0	0	- .4	- .4
10	0	+ 1	+ .2	0	0	- .1	- .1
11	- 1	- 3	- .6	0	0	+ .3	+ .3
12	- 2	- 4	- .8	0	0	+ .4	+ .4
13	- 2	- 2	- .4	0	0	+ .2	+ .2
14	+ 1	+ 1	+ .2	0	0	- .1	- .1
15	+ 2	+ 3	+ .6	0	0	- .3	- .3
16	+ 2	+ 4	+ .8	0	0	- .4	- .4
17	0	+ 1	+ .2	0	0		
18	- 1	- 3	- .6				
19	- 2						
20	- 2						

Diagram 34 shows the original series with the results of using 5-, 7-, and 9-moving averages. The original series consists of regular fluctuations repeating themselves at the end of 7-time intervals. The moving averages in groups of 7 absolutely eliminates the fluctuations. The 5- and 9-moving averages leave us with series which still contain fluctuations, although the extent of these has been much reduced. In the original series the fluctuations range between  $\pm 2$ , the reduced fluctuations in the series of averages of 5 items range between  $\pm .8$ , and those in the series of averages of 9 items between  $\pm .4$ . Further, the series of averages of 9 items have maxima when the original series had minima and vice-versa; we may say that the process of smoothing the fluctuations in the original series has gone too far, the fluctuations have not



only been smoothed but have been reintroduced in the opposite sense to that in which they were, in the original series

This example illustrates certain general principles. If

Time Interval	Series	Sums in 5's	5-Interval Moving Average	Sums in 9's	9-Interval Moving Average	Sums in 15's	15-Interval Moving Average
1	-2						
2	0						
3	+1	-1	-.2				
4	0	-2	-.4				
5	0	-1	-.2	-5	-.6		
6	-1	-1	-.2	-4	-.4		
7	-1	-4	-.8	-5	-.6		
8	+1	-5	-1.0	-5	-.6	-4	-.3
9	-3	-5	-1.0	-2	-.2	-3	-.2
10	-1	-3	-.6	-3	-.3	-3	-.2
11	-1	-1	-.2	-2	-.2	-3	-.2
12	+1	+1	+.2	-2	-.2	-1	-.1
13	+3	+2	+.4	-3	-.3	-2	-.1
14	-1	+2	+.4	+1	+.1	-1	-.1
15	0	+1	+.2	+4	+.4	+2	+.1
16	-1	-1	-.2	+4	+.4	+2	+.1
17	0	+2	+.4	+3	+.3	+5	+.3
18	+1	+1	+.2	+2	+.2	+8	+.5
19	+2	+2	+.4	+4	+.4	+7	+.5
20	-1	+4	+.8	+4	+.4	+3	+.2
21	0	+4	+.8	+7	+.8	+2	+.1
22	+2	+2	+.4	+5	+.6	+4	+.3
23	+1	+5	+1.0	+1	+.1	+2	+.1
24	0	+3	+.6	+1	+.1		
25	+2	-2	-.4	+3	+.3		
26	-2	-1	-.2	+1	+.1		
27	-3	0	0				
28	+2	-4	-.8				
29	+1						
30	-2						

the regular fluctuations exactly repeat themselves at the end of  $n$  time intervals, then an  $n$ -time interval moving average will eliminate the fluctuations entirely. If the process of moving averages is used, taking groups of less

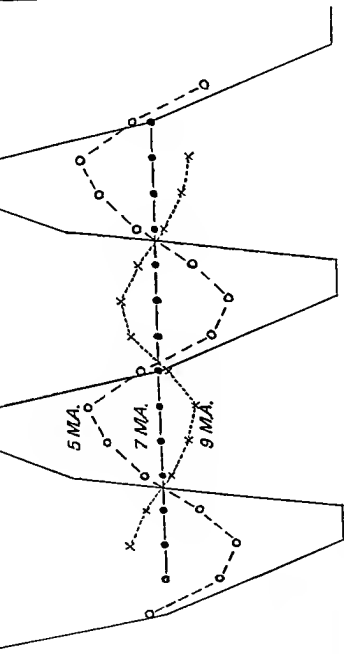


DIAGRAM 34 Series with 5-, 7-, and 9-Interval Moving Average

than  $n$  items, the fluctuations are merely reduced but not eliminated, if more than  $n$  items but less than  $2n$  are taken in the averaging process the fluctuations are still further reduced, but now the result gives a series with maxima when the original series had minima and vice-versa. If  $2n$  items are taken in the averaging process the fluctuations are again eliminated, and so on. Thus, to eliminate entirely the fluctuations, we must use the method of moving averages with  $n$ ,  $2n$ ,  $3n$ , etc., items.

Let us now consider the effect of using the method of moving averages on a series of irregular fluctuations. We can illustrate with the series above (table on p 222), on which the moving average process has been used, taking different numbers of items together.

The results of using the process on this series are apparent. The size of the fluctuations is reduced, but the fluctuations are not eliminated. The original fluctuations range between  $\pm 3$ , the 5-, 9-, 15-interval moving averages range between  $\pm 1$ ,  $+.8$  and  $-.6$ ,  $+.5$  and  $-.3$  respectively. Thus with a larger number of items in the average the extent of the fluctuations is greatly reduced. This illustrates a general principle. With adventitious fluctuations, positive and negative signs occurring at random, the more items taken together the more chance there is of the positive and negative amounts occurring in equal numbers in any group, and, therefore, cancelling to a large extent. But we should never expect always to get complete cancellation, the successive sums would be generally different from zero, but small. The larger the number of items in the group the larger the denominator in the average and, therefore, the smaller the average.

Thus the most complete reduction in the extent of the

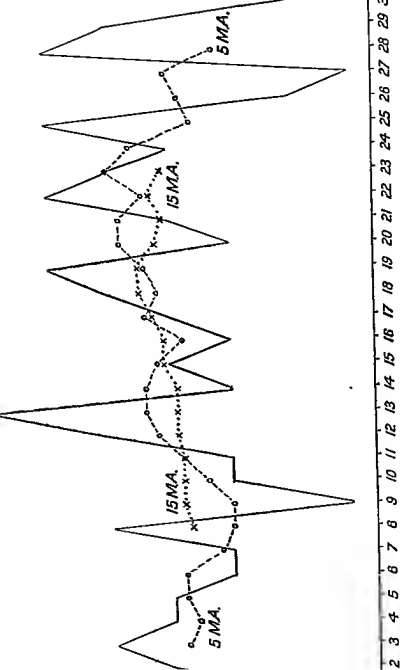


DIAGRAM 35 Series with 5- and 15-Interval Moving Average

fluctuations will be obtained when we take as large a number as possible, of items, in our smoothing process.

We may summarize now the results so far obtained of using the moving average method.

Trend	Linear	Reproduces the trend exactly
	Curved	Reduces the curvature, the more items we take in the moving average, the more remote is the result from the original
Fluctuations	Regular	That number of items in the average which agrees with the period of the fluctuations eliminates them entirely, or any multiple of that number. Other groupings of items merely reduce the extent of the fluctuations
Fluctuations	Irregular	These are never completely eliminated, but they are considerably reduced, and the greater the number of items in the average the more is the reduction in the fluctuations

When we wish to combine these results into a working rule for guidance in using the method on a future occasion, we find ourselves on the horns of a dilemma. If we use too many items in our averages we shall do well, as far as the irregular fluctuations are concerned, but we may distort the trend and may not properly eliminate the regular fluctuations. If we use too few, we may not reduce sufficiently the irregular fluctuations, we may not get rid of the regular fluctuations, but we probably will not introduce distortion into the trend. In practice, we adopt a middle course, and take as the number of items in the averaging process that which will eliminate the regular fluctuations, hoping that we shall thus reduce very considerably the random fluctuations without introducing too much distortion into the trend. Thus, in practice, when we wish to use the moving average method we search for periodicity in our series. If, from a diagram, we estimate that an annual series (say) appears to have a regular up

and down movement repeated at intervals of seven years, we use a 7-year moving average in order to smooth out the fluctuations, and hope that the result will give us a very good approximation to the trend.

Let us consider the series which was graphically represented in Diagram 32. As far as we can judge from

Year	Series	5-year Moving Average	8-year Moving Average	9-year Moving Average
1	11.1			
2	13.1			
3	8.3	10.5		
4	10.0	10.1		
5	10.1	9.8	10.6	10.7
6	9.2	10.4	10.7	10.8
7	11.4	10.8	10.6	10.8
8	11.5	11.2	10.9	10.7
9	11.7	11.6	10.8	10.7
10	12.3	11.2	10.8	10.8
11	11.0	10.9	10.9	11.2
12	9.4	10.6	11.1	11.2
13	10.0	10.7	11.1	11.3
14	10.3	10.8	11.2	11.5
15	13.0	11.4	11.4	11.6
16	11.4	12.2	11.8	11.5
17	12.3	12.7	11.7	11.6
18	14.2	12.1	12.0	11.9
19	12.6	12.3	12.1	12.2
20	10.1	12.0	12.1	
21	12.4	11.7		
22	10.7			
23	12.9			

this graph, there are prominent peaks in years 2, 10, 18, at intervals of 8 years, and prominent depressions in years 3, 12, 20, at intervals of 9 and 8 years. We can reduce the irregular fluctuations by using (say) a 5-year moving average. This would also have the effect of showing us if the regular movement was repeated at 8- or 9-year

intervals. This is done and the results in Diagram 36 show peaks at 9 and 17 years and depressions at 5 and 12 or 13 years. Let us then use an 8-year and a 9-year moving average on the series. The results are shown in the table on p. 227.

Diagram 36 also shows the results of the 8-year moving average smoothing. The fluctuations are nearly completely eliminated. It will be observed that the points on the moving average graph are placed in between ordinates drawn through points on the time axis corresponding to the given years. If we wished to estimate the values on the smoothed graph (from the 8-year moving average), corresponding to the original values in the series, we proceed as follows :—

Year	Series	Sum in 8's	Add in pairs	Divide by 18	Original Trend
1	11.1				
2	13.1				
3	8.3				
4	10.0				
5	10.1	84.7	170 0	10.6	10.4
6	9.2	85.3	169 8	10.8	10.5
7	11.4	84.5	171.7	10.7	10.6
8	11.5	87.2	173.8	10.9	10.7
9	11.7	86.6	172.1	10.8	10.8
10	12.3	86.5	174.1	10.9	10.9
11	11.0	87.6	176.8	11.0	11.0
12	9.4	89.2	178.3	11.1	11.1
13	10.0	89.1	178.8	11.2	11.2
14	10.3	89.7	181.3	11.3	11.3
15	13.0	91.6	184.8	11.5	11.4
16	11.4	93.2	187.1	11.7	11.5
17	12.3	93.9	190.2	11.9	11.6
18	14.2	96.3	193.0	12.1	11.7
19	12.6	96.7	193.3	12.1	11.8
20	10.1	96.6			
21	12.4				
22	10.7				
23	12.9				

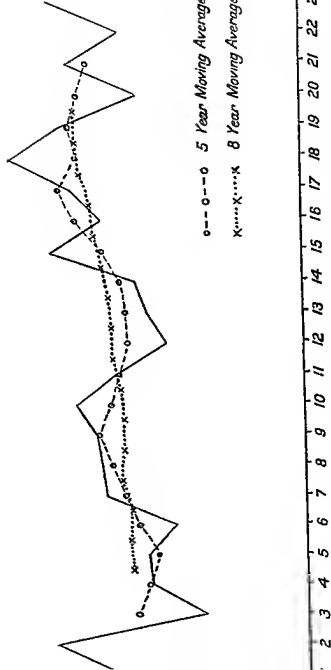


DIAGRAM 36

The figures in the column headed "Divide by 16" are the values on the 8-year smooth curve for the years for which the original figures in the series were given. These smooth values are really the averages of the figures obtained as a result of the 8-year moving average process. It is

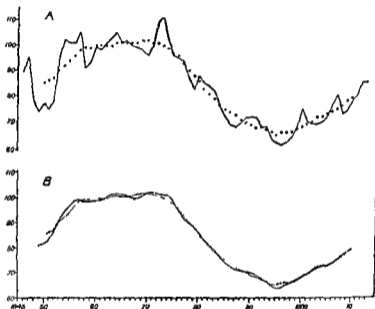


DIAGRAM 37 (A) *Statist Wholesale Price Index* and 9-year moving average (B) 7- and 9-year moving average

interesting to compare the results with the original trend figures from which our series was obtained. These last are shown in the table on p 228 side by side with the smooth series. They differ at most by .4; the difference is, of course, due to the fact that the irregular fluctuations are not completely eliminated.

One disadvantage in the use of this method is that trend values corresponding to a first and last group of

Year	Index Number	7-year Moving Average	9-year Moving Average	Year	Index Number	7-year Moving Average	9-year Moving Average
1846	89			1880	88	86.1	86.0
1847	95			1881	85	83.6	83.4
1848	78			1882	84	81.4	80.7
1849	74	80.9		1883	82	79.4	78.6
1850	77	81.7	84.8	1884	76	76.6	77.1
1851	75	82.7	86.1	1885	72	74.4	75.3
1852	78	86.0	86.8	1886	69	72.7	73.9
1853	95	69.9	89.8	1887	68	71.3	72.8
1854	102	93.9	91.7	1888	70	70.7	71.0
1855	101	96.1	93.6	1889	72	70.1	70.1
1856	101	98.4	99.1	1890	72	70.0	69.1
1857	105	99.0	98.4	1891	72	69.3	68.3
1858	91	98.4	99.2	1892	68	68.1	87.6
1859	94	98.4	99.2	1893	68	68.8	68.7
1860	89	98.7	99.7	1894	63	85.1	85.8
1861	98	93.7	99.7	1895	62	64.0	85.8
1862	101	100.1	99.3	1896	81	64.0	85.7
1863	103	101.3	100.3	1897	62	65.0	85.9
1864	105	101.4	100.9	1898	64	66.0	88.0
1865	101	101.6	100.8	1899	68	87.0	68.7
1866	102	101.1	100.8	1900	75	68.1	87.8
1867	100	100.1	100.4	1901	70	69.3	68.8
1868	99	99.4	101.1	1902	69	70.4	70.4
1869	98	100.6	101.8	1903	69	71.7	72.2
1870	98	101.9	101.9	1904	70	72.4	72.8
1871	100	102.1	101.2	1905	72	72.9	72.7
1872	109	101.7	100.7	1906	77	73.6	73.6
1873	111	101.3	100.1	1907	80	74.9	74.8
1874	102	101.0	98.9	1908	73	76.3	76.6
1875	96	99.1	97.4	1909	74	78.1	78.2
1876	95	95.4	96.1	1910	78	79.3	
1877	94	92.1	93.4	1901	80		
1878	87	89.7	90.4	1912	83		
1879	83	88.0	88.2	1913	85		

values of the series are not obtained. On the other hand, it is a method which is easy of application, and produces results which are the same whoever is operating. Thus

last is a distinct advantage over the freehand curve method, which has the merit of being simpler.

When we are dealing with certain series of economic

Year	Series	Trend	Fluctuation	Year	Series	Trend	Fluctuation
1846	89			1880	88	86	+ 2
1847	95			1881	85	83	+ 2
1848	78			1882	84	81	+ 3
1849	74			1883	82	79	+ 3
1850	77	85	- 8	1884	78	77	- 1
1851	75	86	- 11	1885	72	75	- 3
1852	78	87	- 9	1886	69	74	- 5
1853	95	90	+ 5	1887	68	73	- 5
1854	102	92	+ 10	1888	70	71	- 1
1855	101	94	+ 7	1889	72	70	+ 2
1856	101	96	+ 5	1890	72	69	+ 3
1857	105	98	+ 7	1891	72	68	+ 4
1858	91	99	- 8	1892	68	68	0
1859	94	99	- 5	1893	68	67	+ 1
1860	89	100	- 1	1894	63	66	- 3
1861	98	100	- 2	1895	62	65	- 3
1862	101	100	+ 1	1896	81	66	- 5
1863	103	100	+ 3	1897	62	66	- 4
1864	105	101	+ 4	1898	64	68	- 2
1865	101	101	0	1899	63	67	+ 1
1866	102	101	+ 1	1900	75	68	+ 7
1867	100	101	- 1	1801	70	69	+ 1
1868	99	101	- 2	1902	69	70	- 1
1869	98	102	- 4	1903	69	72	- 3
1870	96	102	- 6	1904	70	72	- 2
1871	100	101	- 1	1905	72	73	- 1
1872	109	101	+ 8	1906	77	74	+ 3
1873	111	100	+ 11	1907	80	75	+ 5
1874	102	99	+ 3	1908	73	76	- 3
1875	96	97	- 1	1909	74	78	- 4
1876	95	96	- 1	1910	78	79	- 1
1877	94	93	+ 1	1911	80		
1878	87	90	- 3	1912	85		
1879	83	88	- 5	1913	85		

data, although we may see that fairly regular fluctuations exist, we may have some doubt as to the exact periodicity of the movement; thus, if we consider a graph of the

*Statist* Wholesale Price Index from 1846-1913, Diagram 37, we can see an up and down movement repeated more or less every seven years. Thus there are peaks in 1900 and 1907, in 1873 and 1880, in 1857 and 1864. But the interval between 1864 and 1873 is nine years. There are depressions in 1858 and 1870 (12 years' interval) and 1879 (9 year's interval) and 1887 (8 year's interval) and 1896 (9 year's interval). The movement is not perfectly regular, and if we use a 9-year moving average it will give satisfactory results in some parts of the range, but in others a 7-year moving average might be preferable. Let us consider the results of operating both methods in the series (table on p. 231).

The results of using the moving average process are shown in Diagram 37. There are still apparent some slight perturbations in the 7-year graph, and up to 1898 it is probably best to take the results of the 9-year smoothing as the trend, and after that the 7-year smoothing. We may even smooth out any slight irregularity remaining in the results of using moving averages, e.g. in 1862 and 1867, and obtain the series on p. 232 as the trend. The table shows this together with the original series and the fluctuations from the trend.

Let us now consider the problem of obtaining the extent of the regular movement, when it exists, in a series. We can illustrate this by dealing with a series possessing a seasonal movement. Such a series is that below, giving output of coal in Great Britain, quarterly, which we would anticipate will exhibit a seasonal movement on account of the changing demand for coal with the seasons. We operate on the series with a 4-quarter moving average, in order to obtain the trend. Knowing this we can get the fluctuations

from it, and analyse these in order to arrive at the seasonal movement.

GREAT BRITAIN: QUARTERLY OUTPUT OF COAL (MN. TONS)

	Quarters	Output	Sums in 4's	Add in pairs	Divide by 8 (Trend)	Fluctua- tions from the Trend
1927	1	68.3				
	2	62.6				
	3	61.1	255.3	507.7	63.5	- 2.4
	4	63.3	252.4	500.1	62.5	+ 0.8
1928	1	65.4	247.7	490.7	61.3	+ 4.1
	2	57.9	243.0	484.2	60.5	- 2.6
	3	56.4	241.2	485.1	60.8	- 4.2
	4	61.5	243.9	492.6	61.6	- 0.1
1929	1	68.1	248.7	503.8	63.0	+ 5.1
	2	62.7	255.1	515.7	64.5	- 1.8
	3	62.8	260.6	523.2	65.4	- 2.8
	4	67.0	262.8	521.6	65.2	+ 1.8
1930	1	70.1	259.0	511.5	63.9	+ 6.2
	2	59.1	252.5	499.6	62.4	- 3.3
	3	56.3	247.1	483.6	60.4	- 4.1
	4	61.8	236.5	468.7	58.6	+ 3.0
1931	1	59.5	232.2	459.2	57.4	+ 2.1
	2	54.8	227.0	450.4	56.3	- 1.5
	3	51.1	223.4			
	4	58.0				

This series of fluctuations consists of both regular and irregular perturbations, the periodicity of the regular movement being four quarters. We arrange the fluctuations as follows :—

FLUCTUATIONS FROM THE TREND

Quarters	1	2	3	4
1927			- 2.4	+ 0.8
1928	+ 4.1	- 2.6	- 4.2	- 0.1
1929	+ 5.1	- 1.8	- 2.6	+ 1.8
1930	+ 6.2	- 3.3	- 4.1	+ 3.0
1931	+ 2.1	- 1.5		
Totals	+ 17.5	- 9.2	- 13.3	+ 5.5
Averages	+ 4.4	- 2.3	- 3.3	+ 1.4

The last line gives the regular seasonal movement.

Each first quarter's figure is the result of the regular seasonal influences plus other adventitious influences which do not operate in a regular periodic fashion. We assume with respect to the latter, that over a long time their influence will sometimes be to increase the normal seasonal movement and sometimes to diminish it, and the results of these factors will more or less cancel when we

	Quarter	Trend	Seasonal Move- ment	Irregular Fluctua- tions	Original Series
1927	3	63.5	- 3.3	+ 0.9	61.1
	4	62.5	+ 1.4	- 0.6	63.3
1928	1	61.3	+ 4.4	- 0.3	65.4
	2	60.5	- 2.3	- 0.3	57.9
	3	60.6	- 3.3	- 0.9	56.4
	4	61.6	+ 1.4	- 1.5	61.5
1929	1	63.0	+ 4.4	+ 0.7	68.1
	2	64.5	- 2.3	+ 0.5	62.7
	3	65.4	- 3.3	+ 0.7	62.8
	4	65.2	+ 1.4	+ 0.4	67.0
1930	1	63.9	+ 4.4	+ 1.8	70.1
	2	62.4	- 2.3	- 1.0	59.1
	3	60.4	- 3.3	- 0.6	56.3
	4	58.6	+ 1.4	+ 1.6	61.6
1931	1	57.4	+ 4.4	- 2.3	59.5
	2	56.3	- 2.3	+ 0.8	54.8

add them together. Naturally the more years' figures we can take, the more likely is this cancellation to be complete and, at any rate, the average of this random part of the fluctuations is likely to be very close to zero. So we assume that the average of the first quarter's figures over a period of years represents only the regular movement, that, by averaging, the random fluctuations have

been eliminated. Similarly we do likewise with the other quarters' figures, and assume that the result gives only the regular movement. In this kind of calculation

MONTHLY COST OF LIVING INDEX NUMBERS OF THE MINISTRY OF LABOUR

Months	Index	12-month M A Trend	Fluctua- tions	Months	Index	12-month M A. Trend	Fluctua- tions
1927, 1	175			1930, 1	166	161.7	+4.3
2	172			2	164	161.2	+2.8
3	171			3	161	160.7	+0.3
4	165			4	157	160.0	-3.0
5	164			5	155	159.3	-4.3
6	163			6	154	158.3	-4.3
7	166	167.2	-1.2	7	155	157.3	-2.3
8	164	166.7	-2.7	8	157	156.2	+0.8
9	165	166.1	-1.1	9	157	155.3	+1.7
10	167	165.8	+1.2	10	156	154.4	+1.6
11	169	165.8	+3.2	11	157	153.7	+3.3
12	169	165.8	+3.2	12	155	153.0	+2.0
1928, 1	168	165.9	+2.1	1931, 1	153	152.2	+0.8
2	168	165.9	+0.1	2	152	151.4	+0.8
3	164	165.9	-1.9	3	150	150.4	-0.4
4	164	165.9	-1.9	4	147	149.5	-2.5
5	154	165.8	-1.8	5	147	148.5	-1.5
6	165	163.6	-0.6	6	145	147.8	-2.8
7	165	165.5	-0.5	7	147	147.2	-0.2
8	165	163.5	-0.5	8	145	146.8	-1.8
9	165	163.5	-0.5	9	145	146.4	-1.4
10	166	165.5	+0.5	10	145	146.1	-1.1
11	167	163.3	+1.7	11	146	145.8	+0.2
12	168	165.0	+3.0	12	148	145.5	+2.5
1929, 1	167	164.6	+2.4	1932, 1	147	145.2	+1.8
2	165	164.3	+0.7	2	147	144.9	+2.1
3	166	164.2	+1.8	3	146	144.6	+1.4
4	162	164.1	-2.1	4	144	144.3	-0.3
5	161	164.1	-3.1	5	143	144.1	-1.1
6	160	164.0	-4.0	6	142	143.8	-1.8
7	161	164.0	-3.0	7	143		
8	163	163.9	-0.9	8	141		
9	164	163.6	+0.4	9	141		
10	165	163.2	+1.8	10	143		
11	167	162.8	+4.2	11	143		
12	167	162.2	+4.8	12	143		

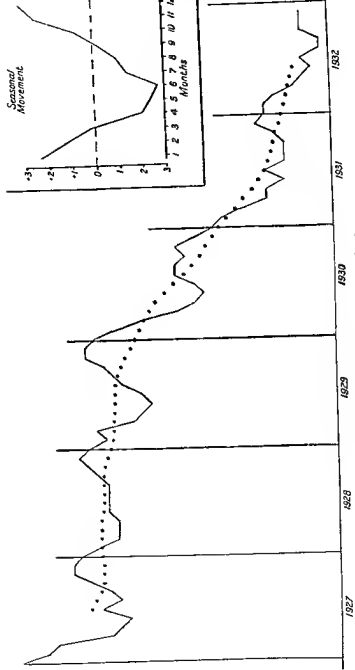


DIAGRAM 38 Cost of Living Index Number

we take as many years' experience as is available, in order to reduce as much as possible the random element left in the fluctuations after the trend has been taken away from the original series. In the simple illustration shown here there is only the experience of four years from which to estimate the regular movement, and we should not, in practice, expect the results to be very precise.

We will suppose that the results obtained do actually give us the seasonal movement, then we can get the irregular movement and present the final results of the analysis as in the table on p. 235.

This method is applied in exactly the same kind of way to a series of monthly figures. We may illustrate with the monthly cost of Living Index Numbers of the Ministry of Labour. (Table on p. 236.)

FLUCTUATIONS CALCULATION OF REGULAR SEASONAL MOVEMENT

Month	1927	1928	1929	1930	1931	1932	Totals	Regular Seasonal Movement
1		+ 21	+ 24	+ 43	+ 08	+ 18	+ 11.4	+ 23
2		+ 01	+ 07	+ 28	+ 06	+ 21	+ 63	+ 13
3		- 19	+ 18	+ 03	- 04	+ 14	+ 12	+ 02
4		- 19	- 21	- 30	- 25	- 03	- 98	- 20
5		- 18	- 31	- 43	- 15	- 1.1	- 118	- 24
6		- 06	- 40	- 43	- 28	- 18	- 135	- 27
7	- 12	- 05	- 30	- 23	- 02		- 7.2	- 14
8	- 27	- 05	- 09	+ 08	- 18		- 51	- 10
9	- 11	- 05	+ 04	+ 17	- 14		- 09	- 02
10	+ 12	+ 05	+ 18	+ 16	- 11		+ 40	+ 08
11	+ 32	+ 17	+ 42	+ 33	+ 02		+ 126	+ 25
12	+ 32	+ 30	+ 48	+ 20	+ 25		+ 155	+ 31

There is, in the Cost of Living Index Number, a seasonal movement which raises and lowers the number by as much

as 3 points, the maximum increase being in December, and the maximum decrease in June. This seasonal movement is in the main due to higher prices in winter of certain foodstuffs.

Diagram 38 shows the change in the Cost of Living Index Number.

# INDEX

*The numbers refer to pages*

- Abstract, Statistical, 37
- Accuracy of Averages and Ratios, 51-7; of Statistical Data, 10, 11, 24
- Acreage under Crops, 184, 206-7
- Ages: at Death, 38; at Marriage, 34, 102
- Age-distribution: of Female Textile Workers, 101, of Population, 43
- Area Scale, in graphs, 99, when grade intervals change, 102
- Assembly of Statistical Data, 24
- Average: Estimate of, 84-94; Limitations of, as representative of a group, 120; Persons per Room, 82; Persons per Family, 79; Rooms per Family, 81; Wagon Load, 65, 120; Weighted, 141; Weighted, Difference between weighted and unweighted, 143
- Averages, 42, 48-50, Calculation of, 77, Comparison of, 60; Utility of, 74, Moving, *see* Moving Averages
- Balance of Trade, 57
- Birth-rate, 64, 182
- Births, Number of, per Marriage, 77
- Board of Trade, 154, 158, 168  
*See also* Index Number of Wholesale Prices
- Bowley, Prof A L. 9, 16
- Calculation of Averages, 77, Mean Deviation, 123, Quartiles and Quartile Deviation, 132; Standard Deviation, 126
- Census, 14, 18; of Population, 8, 11, 14, 23, 25, 29, 40, 44, 56, 78, 176; of Production, 8, 14
- Chain Base in Index Numbers, 156
- Characteristics of Units in a Statistical Inquiry, 22, 25
- Classification, 25-8
- Coal: Mining Industry, 33, 37, 55, 60, 70, 75, 91, 100, 119, 180, 189, 200; Production, 177, 208, 234
- Comparisons, 42; of Averages, 60
- Cumulative: Diagrams, 108; Tables, 109
- Cyclical Movement in Time Series, 210
- Data, Statistical, 2-3; Precise meaning of, 3-4, 9
- Death-rate, 62, 73; as simple or weighted average, 148
- Definitions, 4, 8, 9; and Classification, 28
- Deviation, 122; Average or Mean, 122, Mean, from Average or Median, calculation of, 123; Quartile, 131, Standard, 126
- Diagrams, 97, 173; Cumulative, 108; showing grouped data (linear and area scale), 88, 99
- Distribution of Unemployment Percentages, 106
- Estimate of Average, 84-94; Insured Workers, 57; Population, 56
- Export Trade, 14, 27, 56, 159
- Foreign Trade, 27, 37, 56, 159

- Geometric Mean, 162-3  
 Graphical Methods, 97  
 Graphs with linear and area scale, 98-9  
 Grouping Statistical Data, 27  
 Histogram, 105  
 Imports, 7, 8, 10, 14, 19, 27, 56, 159, 179, 183  
 Index of Production, 151, 154, 155, 158, 160  
 Index Number, 151, Cost of Living, 153, 157, 167, 168, 170, Fixed Base, 155, Moving Base, 155; of Prices, 152, 155, of Wholesale Prices, Board of Trade, 158, 164, of Wholesale Prices, Statist, 155, 156, 161, 164, 230-3, weights, 159  
 Industry, Numbers in, 47  
 Inquiries, Statistical, Kinds of, 7, 8  
 Insured Population, 57, 173, 176  
 Interpretation of Ratios, 62  
*Labour Gazette* 154  
 Linear Scale in Graphs, 96  
 Logarithm Tables, 57  
 Logarithms used in Calculation of Geometric Mean, 162  
 Logarithmic Scale, 196  
 London and Cambridge Economic Service, 139, 155, 158  
 Marks in an Examination, 84, 99  
 Marriage-rate, 63  
 Mean *same as simple average*, see average, Deviation, calculation of 123, Geometric, 162  
 Measures of dispersion in a group, 121-134  
 Median, 116, 117, by Calculation, 118, from Cumulative Diagram, 118  
 Medical Inspections, School, 12  
 Methods, Statistical, 24  
 Mineral Output, 185  
 Mines, Annual Report of Secretary for, 33  
*Ministry of Labour Gazette*, 154  
 Moving Averages, 214, Effect of using the method on a series without fluctuations 216, Effect of using the method on a series of irregular fluctuations, 222, Effect of using the method on a series of regular fluctuations, 219, General guidance for using the method, 226  
 Output Gross, 181, Net 160, of Coal, 177, 208, 234, of Minerals, 185, per Man-shift in the Coal Mining Industry, 70, 149, 189  
 Percentages, 48  
 Percentage Scale in diagrams of Time Series, 192  
 Polygon of Frequency, 104  
 Population, 43, 44, 176  
 Price Index Number, 152, 155, 158, 161, 164  
 Primary Statistics, 24  
 Production, Index Number, 151, 154, 155, 158, 160  
 Profits in Coal Mining Industry 91  
 Quartiles Lower and Upper, 131, Lower and Upper from Cumulative Diagram and from Cumulative Table 132  
 Quartile Deviation 131  
 Railway Statistics, 65  
 Range of Variation as measure of dispersion in a group, 121  
 Rate Death, 62, 73, as simple or weighted average 148  
 Rates, 50, Standardized, 67-74  
 Ratio Scales, 196  
 Ratios Statistical 42, Utility of, 74  
 Rectangles used in Graphs, 99, 174

- Regular Movement in a Time Series, 233
- Round Numbers, 57
- Rowe, J. W. F., 139, 155
- Sample, 14, 157, 159; Inquiry, 16; Tests of Random nature of, 18, 58
- Scales used in plotting Time Series, 189, 192, 196
- School Medical Inspections, 12
- Seasonal Movement in Time Series, 210, 233, 238
- Secondary Statistics, 24, 42
- Secular Trend, *see* Trend
- Series, Time, 173
- Shipping, 37
- Significant Figures, 53
- Snow, Dr. E. C., 137
- Sources of Statistical Data, 11, 12
- Standard Deviation, 126
- Statistical: Abstract, 37; Aspect of a problem, 5; Data, 2, 3; Data, precise meaning of, 3; 4; Inquiries, 7; Methods, 24; Tables, 24
- Statistician's functions, 1, 2
- Statist, Index Number of Wholesale Prices, 155, 158, 161, 164, 230, 231, 233
- Sum, Weighted, 136
- Super Tax, Numbers liable to, 94, 109, 203
- Survey, London (*New Survey of London Life and Labour*, 1928), 9, 20, 137
- Tables: Cumulative, 109; Statistical, 24, 32-40
- Tabulation, 24
- Tests of Random Sampling, 18, 58
- Time Series, 173; Trend, 209, 214; Seasonal Movement, 210, 233; Cyclical Movement, 210; Irregular fluctuations, 211
- Trade: Balance of, 57; Board of, 154, 158; *see also* Board of Trade and Index Number of Prices
- Trapezia, used in graphs, 104
- Trend in Time Series, 209
- Unemployed Person, 8, 9, 20, 57
- Unemployment: Insurance, 7, 17; Percentages, Cumulative table, 110, 112; Percentages, distribution by districts, 106
- Units in Statistical Inquiry, 22
- Variation in a Group, 120
- Weighted: Averages, 141; Averages, difference between weighted and unweighted average, 143; Sums, 136
- Weights in Index Number Calculations, 159
- Wheat consigned to U.K., 187
- Working Class, 9, 20